U N I V E R Z A V L J U B L J A N I
Fakulteta za elektrotehniko

Janez Žibert

OBDELAVA IN ANALIZA ZVOČNIH POSNETKOV INFORMATIVNIH ODDAJ Z UPORABO GOVORNIH TEHNOLOGIJ

Doktorska disertacija

Mentor: prof. dr. France Mihelič

Ljubljana, 2006

Moji družini.

Zahvala

Doktorska disertacija je plod mojega raziskovalnega dela v Laboratoriju za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko Univerze v Ljubljani. Zato bi se ob tej priložnosti zahvalil vsem sodelavcem laboratorija, ki so mi omogočili vse potrebne pogoje za ustvarjalno znanstveno–raziskovalno okolje, v katerem je nastajala moja disertacija.

V prvi vrsti bi se rad zahvalil svojemu mentorju prof. dr. Francetu Miheliču, ki me je skozi celotni podiplomski študij usmerjal in mi svetoval pri delu. Njegova neizmerna potrpežljivost ter pripravljenost pomagati pri razvijanju in udejanjanju novih idej sta ključno prispevali k nastanku disertacije. Še posebej bi se mu rad zahvalil tudi za to, da mi je v zadnjem letu nastajanja disertacije odstopil del svojega kabineta na fakulteti, kjer sem lahko našel ustvarjalni mir za pisanje disertacije.

Zahvala gre tudi prof. dr. Nikoli Pavešiću, vodji laboratorija, ter vsem sodelavcem laboratorija za vso podporo pri mojem raziskovalnem delu. Še posebej bi se rad zahvalil Simonu, ki me je že v začetku moje raziskovalne poti navdušil s svojim poglobljenim znanjem s področja govornih tehnologij in je bil vedno pripravljen deliti svoje ideje z mano. Izpostavil bi tudi Boštjana, ki mi je s svojimi konstruktivnimi komentarji in včasih tudi z drugačnimi pogledi na probleme, s katerimi sem se ukvarjal, zelo pomagal pri raziskovalnem delu. Seveda pa se moram zahvaliti tudi ostalim sodelavcem laboratorija, Ankici, Meliti, Jerneji, Tonetu, Ivu, Mariu, Vakilu in Jaki, ki so bili v različnih obdobjih in na različne načine prisotni na moji raziskovalni poti, predvsem pa so ustvarjali prijetno delovno vzdušje.

Osnovne ideje in motivacije za nastanek doktorske disertacije sem pridobil na gostovanju v Lizboni, kjer smo v okviru projekta COST278 na inštitutu *INESC-ID Lisboa* začeli s pridobivanjem in označevanjem posnetkov informativnih oddaj ter definirali probleme, s katerimi sem se ukvarjal v disertaciji. Zato bi se rad zahvalil vsem kolegom iz projekta, s katerimi sem sodeloval pri pripravi in vrednotenju postopkov na zbirki COST278. Pri označevanju in dokumentiranju zbirke SiBN pa bi se rad zahvalil tudi vsem študentom, ki so sodelovali pri označevanju zvočnih posnetkov informativnih oddaj, in Gregi za pripravo jezikovnega korpusa te zbirke.

Posebna zahvala gre staršem in ženi Mateji, ki so me vedno, ne le v času študija, vzpodbujali in mi bili na voljo, ko sem jih potreboval. Brez njih te disertacije nikoli ne bi bilo.

Ključne besede: segmentacija in razvrščanje segmentov zvočnih posnetkov po govorcih, detekcija govora, segmentacija zvočnih posnetkov na govorne in ne–govorne dele, razvrščanje segmentov na govor in ne–govor, segmentacija zvočnih posnetkov glede na zamenjave govorcev, segmentacija zvočnih posnetkov glede na akustične spremembe, razvrščanje segmentov z rojenjem, rojenje z združevanjem, prozodične značilke, podatkovne zbirke zvočnih posnetkov informativnih oddaj, samodejno podnaslavljanje informativnih oddaj, samodejno razpoznavanje govora, samodejna indeksacija zvočnih posnetkov, razpoznavanje govorcev, sledenje govorcev v zvočnih posnetkih

Povzetek

V zadnjem času se obseg in vsebina informacij, podanih v multimedijski obliki, neprestano povečujeta. Zaradi tega običajni postopki za pridobivanje informacij iz podatkov, ki so podani le v tekstovni obliki, ne zadoščajo več in jih je potrebno posplošiti tako, da so primerni tudi za vsebine, podane v drugačnih oblikah. Velik del informativnih vsebin radijskih in televizijskih oddaj predstavljajo zvočni podatki, ki se v veliki meri manifestirajo kot govor. Za pridobivanje informacij, ki jih vsebujejo tovrstni podatki, lahko poleg običajnih postopkov obdelave signalov uporabljamo tudi postopke govornih tehnologij. V doktorski disertaciji smo se tako osredotočili predvsem na postopke priprave in organiziranja zvočnih posnetkov informativnih oddaj, da bi bili primerni za nadaljnjo obdelavo v različnih sistemih pridobivanja informacij z uporabo govornih tehnologij.

Ukvarjali smo se s tremi nalogami obdelave zvočnih posnetkov: z detekcijo govora v zvočnih posnetkih ter s segmentacijo zvočnih posnetkov glede na zamenjave govorcev in spremembe akustičnega ozadja ter z razvrščanjem segmentov po govorcih. Osnovni cilj je bilo izboljšanje obstoječih in razvoj novih postopkov, ki bi jih lahko vključevali v različne sisteme govornih tehnologij. Pri tem smo se ukvarjali predvsem z različnimi predstavitvami zvočnih posnetkov, s katerimi bi bolje opisali lastnosti v signalih, ki smo jih želeli modelirati, in z zanesljivostjo delovanja posameznih postopkov v različnih akustičnih razmerah. Osnovno vodilo pri razvoju postopkov je bila izgradnja sistema za samodejno indeksacijo zvočnih posnetkov po govorcih.

Zvočne posnetke informativnih oddaj, ki smo jih uporabljali za razvoj in vrednotenje postopkov iz disertacije, smo pridobili iz dveh podatkovnih zbirk. V okviru raziskovalnega dela disertacije je bila na novo pridobljena slovenska zbirka posnetkov informativnih oddaj SiBN, uporabljali pa smo tudi zbirko COST278, ki je bila sestavljena iz posnetkov informativnih oddaj v različnih evropskih jezikih. Zbirka SiBN je bila zasnovana za namene izgradnje različnih sistemov za samodejno razpoznavanje govora, ki bodo namenjeni razpoznavanju večjih besedišč govora. Poleg posnetkov informativnih oddaj smo pridobili in ustrezno pripravili tudi jezikovni korpus informativnih oddaj, ki je pridružen osnovni zbirki SiBN. Večjezična jezikovna zbirka informativnih oddaj COST278 je bila pridobljena v okviru mednarodnega sodelovanja v projektu COST278[1]. Sestavljena je iz usklajeno označenih zvočnih posnetkov informativnih oddaj v devetih evropskih jezikih in je primerna za razvoj in vrednotenje postopkov govornih aplikacij, ki so neodvisne od jezika.

V drugem poglavju smo se zato ukvarjali s postopki pridobivanja, označevanja in usklajevanja zvočnih posnetkov informativnih oddaj. Označevanje takšnih posnetkov je zaradi narave informativnih oddaj dokaj zahtevno in zamudno opravilo, saj so za ra-

[1]European COoperation in the field of Scientific and Technical research, COST action 278: Spoken Language Interaction in Telecommunication; trajanje projekta: 2002-2005

zliko od vnaprej načrtovanih govornih zbirk posnetki informativnih oddaj pridobljeni v povsem nenadzorovanem okolju. Glavne značilnosti takšnih posnetkov so, da so pridobljeni v različnih akustičnih razmerah, vsebujejo veliko število govorcev, različne načine govora in različne ne–govorne pojave. V obeh zbirkah je bilo zato potrebno označevati številne govorne in ne–govorne elemente, jezikovne in ne–jezikovne informacije, kvaliteto govora in zvočnih posnetkov, lastnosti govorcev, pridobiti in razvrstiti vsebine novic informativnih oddaj, pravilno postavljati meje med stavki, odseki govorcev in meje med različnimi vsebinami ipd. Pri tem smo se zgledovali po pravilih, ki jih je predstavilo združenje LDC[2] in so jih uporabljali pri označevanju podobnih zbirk v drugih jezikih. Tako smo označili 34 ur posnetkov zbirke SiBN, v zbirki COST278 pa je bilo potrebno uskladiti transkripcije približno 30 ur zvočnih posnetkov.

Osnovna naloga detekcije govornih delov v zvočnih posnetkih je razdeliti zvočne posnetke na dele, ki pripadajo govoru, in na dele, kjer govora ni. V tretjem poglavju smo se zato ukvarjali s postopki segmentacije na govorne in ne–govorne odseke. Tu smo si zastavili dve nalogi: poiskati primerne predstavitve zvočnih signalov za ločevanje govora od ne–govora in vključevanje teh predstavitev v različne postopke segmentacije. Tako smo razvili postopek pridobivanja fonetičnih značilk na podlagi zaporedij razpoznanih osnovnih govornih enot, ki smo jih pridobili s preprostimi sistemi razpoznavanja glasov iz zvočnih posnetkov. Izpeljali smo štiri značilke, ki so temeljile na trajanju in spremembah dveh skupin govornih enot: parov samoglasnik–soglasnik (CVS značilke) ter zvenečih in nezvenečih glasov (VUS značilke). Z uvedbo širših skupin glasovnih enot smo se znebili vplivov delovanja različnih razpoznavalnikov glasov in odvisnosti od jezika razpoznavanja. Osnovni namen vpeljave fonetičnih značilk je bil predvsem v tem, da smo hoteli osnovnim akustičnim predstavitvam zvočnih signalov dodati informacijo višjega reda, ki bi bila manj občutljiva na različne akustične spremembe v zvočnih posnetkih in bolj primerna za ločevanje govora od ne–govora. Fonetične značilke smo primerjali z akustičnimi značilkami koeficientov melodičnega kepstra (MFCC) in z značilkami, ki so bile izpeljane na podlagi mer entropije in dinamizma iz osnovnih sistemov za razpoznavanje glasov. Pri tem smo razvili dva postopka segmentacije: postopek, kjer sta se izvajali segmentacija in razvrščanje hkrati in postopek, kjer se je najprej izvajala segmentacija, nato pa razvrščanje segmentov na govor in ne–govor. V obeh primerih smo za razvrščanje uporabili modele kombinacije Gaussovih porazdelitev (GMM modele), ki smo jih ocenili vnaprej. Pri prvem postopku smo GMM modele vključili v mrežo prikritih Markovovih modelov (HMM modelov), segmentacija pa je potekala po postopku Viterbijevega dekodiranja. V drugem primeru se je najprej izvajala segmentacija glede na akustične spremembe v zvočnih posnetkih, nato pa razvrščanje z GMM modeli. Ukvarjali smo se tudi z združevanjem različnih predstavitev, ki smo jih izvajali v postopkih segmentacije s fuzijo. V obsežnih preizkusih smo potrdili neobčutljivost in zanesljivost fonetičnih značilk v primerjavi samo z akustičnimi značilkami ter predstavitvami na podlagi entropije in dinamizma. Najboljše rezultate detekcije govora pa smo dosegli s postopkom segmentacije, kjer smo združili osnovne akustične značilke MFCC in predlagane fonetične značilke CVS.

Pri samodejni segmentaciji zvočnih posnetkov po govorcih smo se posvečali postopkom

[2]Linguistic Data Consortium, http://www.ldc.upenn.edu/

razdelitve zvočnih posnetkov na segmente, ki pripadajo enemu govorcu v nespremenjenih akustičnih razmerah. V tem primeru govorimo o segmentaciji glede na zamenjave govorcev in glede na spremembe akustičnega ozadja (segmentacija SAG). Osnovne naloge, ki smo jih tu reševali, so zajemale iskanje primernih predstavitev zvočnih posnetkov za segmentacijo, izvedbo postopkov segmentacije in določanje pragov in kriterijev za odločitve za postavitev mej med posameznimi segmenti.

Standardni postopek segmentacije oziroma iskanja mej med segmenti različnih govorcev in akustičnih ozadij se izvaja na podlagi Bayesovega informacijskega kriterija (kriterij BIC), kjer se na podlagi modelov, ocenjenih iz posameznih segmentov, odločamo za postavitev meje ali ne. Odločitev za mejo predstavlja prag odločitve, ki je implicitno vključen v kriterij BIC in ga je potrebno določiti vnaprej. Izbira ustreznega praga odločitve je bistvenega pomena za uspešno segmentacijo in predstavlja tudi največji problem segmentacije s kriterijem BIC. Izkaže se namreč, da je potrebno prag odločitve vedno znova prilagajati na različne akustične razmere v zvočnih posnetkih, saj so v nasprotnem primeru rezultati segmentacije slabi. Zato smo se v četrtem poglavju osredotočili predvsem na izvedbo postopkov segmentacije, ki bi bili čim manj odvisni od izbire odprtih parametrov postopkov in s tem manj občutljivi na spremembe akustičnih razmer v zvočnih posnetkih. V ta namen smo razvili postopek segmentacije z relativno določenim pragom odločitve, kjer smo združili dva obstoječa postopka segmentacije: standardni postopek s kriterijem BIC in postopek DISTBIC. S postopkom DISTBIC smo v prvi fazi ocenili možne vrednosti kriterija BIC in s tem prag odločitve za meje, s standardnim postopkom v drugi fazi pa smo določili meje med segmenti. Vhodni parameter v postopek tako ni bil več absolutni prag odločitve, ampak relativno določeni prag glede na ocenjene vrednosti kriterija BIC, ki so se spreminjale glede na akustične razmere v zvočnih posnetkih. Možnost ocenjevanja vrednosti kriterija BIC smo izrabili tudi v drugem predlaganem postopku segmentacije, ki je temeljil na združevanju različnih akustičnih predstavitev zvočnih signalov. Na podlagi ocen kriterijev BIC različnih predstavitev smo lahko izvajali normalizacijo ocen posameznih kriterijev in s tem združevanje ocen s postopki fuzije. Na ta način smo tako z združevanjem ločenih predstavitev lahko bolje ocenjevali krajše odseke v zvočnih posnetkih, ki jih v primeru skupnih predstavitev pri standardnem postopku segmentacije slabo modeliramo.

Preizkušanje in primerjava predlaganih postopkov z dvema referenčnima postopkoma je bilo izvedeno na razvojnih in testnih posnetkih zbirk SiBN in COST278. Razvojne posnetke smo uporabljali za določitev vseh odprtih parametrov postopkov glede na optimalne rezultate segmentacije in za primerjavo postopkov na celotnem intervalu operativnih točk posameznih postopkov. V slednjih preizkusih smo lahko potrdili večjo zanesljivost in neobčutljivost predlaganih postopkov v primeru različnih (neoptimalnih) izbir pragov odločitve in ostalih odprtih parametrov. Prav tako smo s predlaganimi postopki dosegli tudi boljše rezultate na obsežnih testnih zbirkah posnetkov v primerjavi z referenčnimi postopki ob optimalni izbiri parametrov.

Združevanje segmentov po govorcih predstavlja zadnjo fazo v procesu segmentacije in razvrščanja segmentov po govorcih *(ang. speaker diarisation)*, kjer je cilj pridobiti in povezati med seboj tiste dele - segmente - zvočnih posnetkov, ki pripadajo istim govorcem. S postopki detekcije govora in segmentacije po govorcih rešujemo prvi del naloge, torej pridobivanje segmentov. S postopki razvrščanja segmentov pa te seg-

mente povezujemo v skupine, ki pripadajo istim govorcem. V našem primeru smo izvajali povezovanje z združevanjem segmentov, pri čemer smo uporabljali postopke hierarhičnega rojenja. Raziskovali smo različne predstavitve govornih segmentov, ki bi bili primerni za združevanje po govorcih, in iskali mere podobnosti (različnosti) za kriterije združevanja ter kriterije zaustavitve postopkov rojenja.

Raziskovalno delo v petem poglavju je usmerjeno k izboljšavam osnovnega postopka rojenja z združevanjem, ki se uporablja za razvrščanje segmentov po govorcih. V prvem delu smo raziskovali osnovne predstavitve segmentov združevanja. Tako smo izvedli alternativen pristop združevanja segmentov z uporabo metod razpoznavanja govorcev. Tu smo osnovne segmente govora predstavili z GMM modeli, ki smo jih izpeljali iz splošnih modelov govora (UBM) ob uporabi MAP adaptacije. Pri tem postopku smo se ukvarjali predvsem z različnimi kriteriji združevanja tako predstavljenih segmentov in predlagali novo mero, ki je temeljila na kriteriju BIC. Povsem drugačen pristop smo izvedli v postopku rojenja z združevanjem akustične in prozodične informacije. Tu smo osnovnim akustičnim predstavitvam segmentov želeli dodati še prozodično informacijo, ki bi bila primerna za združevanje segmentov po govorcih. V ta namen smo izpeljali 10 prozodičnih značilk, ki smo jih pridobivali iz energije signala, ocene osnovnega tona v signalu in na podlagi razpoznanih osnovnih glasovnih enot v signalu. Na ta način smo vpeljali informacijo višjega reda v postopke rojenja, s čimer smo želeli izboljšati združevanje segmentov v primeru, ko se samo na podlagi akustične informacije ali pa zaradi slabih akustičnih razmer ne bi znali pravilno odločati za združevanje med posameznimi roji segmentov. Z vpeljavo dodatne prozodične informacije osnovnim akustičnim značilkam smo morali prilagoditi tudi osnovni postopek rojenja, da bi lahko potekalo združevanje segmentov na podlagi kombinacije obeh predstavitev.

V drugem delu poglavja 5 smo preučevali različne kriterije zaustavitve postopkov rojenja. Osnovni kriterij zaustavitve v standardnih postopkih je običajno določen s pragom zaustavitve, ki ga ocenimo iz razvojnih zbirk zvočnih posnetkov. Takšni kriteriji so seveda odvisni od ujemanja razmer med razvojnimi in testnimi posnetki, kar smo želeli odpraviti s predlaganimi kriteriji. Izpeljali smo dva kriterija. Prvi je temeljil na skupnem kriteriju BIC in je primeren v postopkih rojenja, kjer se za mero združevanja uporablja prav tako kriterij BIC. Drugi kriterij zaustavitve je temeljil na relativni oceni mere DER. Pri tem kriteriju smo potrebovali dva različna postopka združevanja in na podlagi primerjave napak enega postopka z drugim smo ocenili možne točke zaustavitve postopkov rojenja.

Vrednotenje postopkov združevanja smo izvajali z mero DER. Osnovno mero DER smo prilagodili tako, da smo z njo ocenjevali učinkovitost postopkov rojenja na celotnem intervalu združevanja segmentov. Na ta način smo ocenjevali kvaliteto postopkov rojenja neodvisno od kriterijev zaustavitve. Skupne rezultate, ki so predstavljali hkrati tudi končne rezultate procesa segmentacije in združevanja segmentov po govorcih, smo pridobili ob uporabi kriterijev zaustavitve. Preizkusi so bili izvedeni podobno kot pri segmentaciji na zbirkah SiBN in COST278. Izvajali smo dve skupini preizkusov, v prvi smo testirali samo kvaliteto združevanja segmentov ob ročno označenih segmentih zvočnih posnetkov, v drugi pa smo testirali razvrščanje segmentov ob samodejni segmentaciji in detekciji govora. Delovanje postopkov rojenja se je v obeh primerih nekoliko razlikovalo. Skupna ugotovitev je bila, da z dodajanjem informacije, s katero se

osredotočamo bolj na govorčeve lastnosti (prozodične značilke, GMM modeli govorcev) kot pa na splošne akustične lastnosti segmentov, dosežemo boljše rezultate razvrščanja segmentov.

V sklepnem delu doktorske disertacije smo obravnavali možne izvedbe predlaganih postopkov in vključevanje v različne sisteme govornih tehnologij. Tu smo se osredotočili predvsem na zasnovo sistema za samodejno indeksacijo zvočnih posnetkov po govorcih, kjer smo natančnejše opredelili vlogo in pomen postopkov, s katerimi smo se ukvarjali v disertaciji.

Keywords: speaker diarization, speech detection, speech/non–speech discrimination, speech/non–speech segmentation, speaker–change detection, acoustic–change detection, speaker segmentation, audio segmentation, speaker clustering, agglomerative clustering, prosody features, broadcast news speech databases, automatic broadcast news transcription, speech recognition, audio indexing, speaker recognition, speaker tracking

Abstract

These days there is an increasing need to deal with the large amounts of multimedia information resulting from the growing demand to shift content–based information retrieval from text to various multimedia sources. The data provided from television and radio broadcast news (BN) programs are one such source of this information. In our research we focus on the processing and analysis of audio BN data, where the main information source is represented by speech data. The main issues in our work concern the preparation and organization of BN audio data for further processing in information audio–retrieval systems based on speech technologies.

The thesis addresses the problem of structuring the audio data in terms of speakers, i.e., finding the regions in the audio streams that belong to one speaker and joining each region of the same speaker together. The task of organizing the audio data in this way is known as speaker diarization and was first introduced in the NIST project of *Rich Transcription* in *"Who spoke when"* evaluations. The speaker–diarization problem is composed of several tasks. This thesis addresses three of them: speech detection, speaker– and background–change detection, and speaker clustering. The main objectives in our research were to develop new representations of audio data that were more suitable for each task and to improve the accuracy and increase the robustness of standard approaches under various acoustic and environmental conditions. The motivation for the improvement of the existing methods and the development of new procedures for speaker–diarization tasks is the design of a system for the speaker–based audio indexing of BN shows.

For the development and assessment of our approaches we used audio data from two BN speech databases; these are presented in Chapter 2 of the thesis. The first database is a BN speech database in the Slovenian language, named the SiBN database. This database is developed within the research work of the thesis and will serve mainly as a speech database for the development of large vocabulary continuous speech recognition systems (LVCSRs) in the Slovenian language. The other database is called the COST278 BN database and is constructed from BN shows in several European languages. The database was constructed by ten institutions that are collaborating in the European COST278 action on Spoken Language Interaction in Telecommunications. The database comprises BN shows in nine languages and is intended to be used mainly for the development and evaluation of language–independent speech applications. Unlike other speech databases, which are designed for special purposes and collected in controlled environments, the audio data of BN shows represent mainly real–world speech. They possess several different acoustic, speech and language properties, and the annotation process is therefore very difficult. Hence, Chapter 2 describes the process of producing the transcriptions of the audio data of BN programs, presents the tools that were used to transform and adjust the transcriptions with audio and video

data, and provides the basic analysis of the acoustic, speech and language properties of both BN databases. Currently, the SiBN database consists of 34 hours of annotated BN shows from one TV station, and the COST278 BN database consists of 30 hours of BN shows from several different languages and TV stations, meaning that the data in the SiBN database are more homogeneous in terms of acoustic and language properties than the data in the COST278 database. Because of the different properties of the audio data we decided to use the data from both BN databases in all our experiments to obtain a more objective assessment of the proposed methods in the thesis.

Chapter 3 addresses the speech–detection task. The objective in speech detection is to find the segments in the audio streams in which speech can be detected and the segments where there is no speech. Therefore, speech detection can be seen as a speech/non–speech segmentation problem where two tasks have to be accomplished: appropriate segmentation of the data according to speech and non–speech events, and classification of the segments into speech and non-speech. In our research we focus on developing new representations of audio signals that are more suitable for speech/non–speech classification, and developing new segmentation procedures to include these representations. We propose a new, high–level representation of audio signals based on phoneme recognition features. Unlike previous model–based approaches, where speech and non–speech classes were usually modeled by several models, we have developed a representation where just one model per class is used in the segmentation process. For this purpose four measures based on consonant–vowel (CVS features) and voiced–unvoiced (VUS features) pairs obtained from different phoneme speech recognizers are introduced. They are constructed in such a way as to be independent of the recognizer and the language, and are applied in two different segmentation-classification frameworks. In the first case the segmentation and classification are made simultaneously using a network of Gaussian mixture models (GMMs) and in the second case the acoustic segmentation is made prior to the speech/non–speech classification, also using GMMs. While the first segmentation system serves as a baseline system, the second segmentation is more suitable for CVS (VUS) features. Both systems were evaluated on the SiBN and COST278 BN databases. The evaluation results indicate that the proposed phoneme recognition features are better than the standard mel–frequency cepstral coefficients (MFCCs) and posterior probability–based features (entropy and dynamism). The proposed features proved to be more robust and less sensitive to different training and unforeseen conditions. Additional experiments with fusion models based on cepstral and the proposed phoneme recognition features produced the highest scores overall, which indicates that the most suitable method for speech/non–speech segmentation is a combination of low–level acoustic features and high–level recognition features.

Chapter 4 is dedicated to the task of speaker– and (acoustic) background–change detection in audio data. The objective here is to find the points in the audio stream where the change between two different speakers or acoustic environments occurs. These changing points divide the audio stream into homogeneous regions corresponding to one speaker in an unchanged acoustic environment. These regions are called segments and the procedure for obtaining such segments is called speaker–based audio segmentation. Our research was focused on obtaining the proper representations of audio signals for speaker segmentation and improving the existing segmentation methods so

that they would be more robust in different acoustic conditions.

The majority of existing methods for finding change–detection points in audio data are based on Bayesian information criterion (BIC). The main point here is to estimate the probability models (probability distribution) of two neighboring segments and compare them with the BIC. If the estimated BIC score is under the given threshold, a change point is detected. The threshold, which is implicitly included in the penalty term of the BIC, has to be given in advance and estimated from the training data. The accuracy of a segmentation thus heavily depends on properly estimated thresholds, and this represents the main drawback of such segmentation systems. In our research we tried to overcome this problem. Therefore, in Chapter 4 an improved version of the baseline segmentation system is proposed by introducing relative thresholds. These thresholds are estimated continuously from the current acoustic conditions in the audio streams. In the proposed method two approaches are joined: the standard approach with the BIC and the DISTBIC procedure. In the first phase the DISTBIC segmentation is applied to collect all the possible BIC scores. From the BIC scores the threshold for the change–detection points is estimated according to the given relative shift from the estimated maximum BIC values. In the second phase the change–detection points are found by applying the standard BIC–segmentation procedure with a newly estimated threshold. The possibility of estimating the BIC scores in a segmentation is also exploited in the second proposed approach, where we fuse different acoustic representations of audio signals in the segmentation process. The estimation of the BIC scores of the different acoustic representations of audio data allows us to perform the normalization of the BIC scores for each representation, which is used for combining different segmentation procedures in a single fusion system.

We performed several experiments in which we evaluated four different segmentation procedures. The experiments were made on evaluation and test audio files extracted from both BN databases. The evaluation database served for optimal tuning of all the segmentation systems and for a side–by–side comparison of approaches in the domain of different operating points. The last experiments were performed to check the stability and robustness of the proposed and baseline methods in non-optimal working conditions, while the experiments on the test audio files were performed just in the optimal case. In both groups of experiments the proposed methods produced better results than baseline systems, which proves that the proposed methods perform more reliably and stably across different acoustic environments, especially in cases of unmatched training and working conditions.

Speaker clustering represents the last step in the speaker–diarization process. While the aim of speech detection and speaker– and acoustic–segmentation procedures is to provide the proper segmentation of audio data streams, the purpose of speaker clustering is to join or connect together segments that belong to the same speakers. Chapter 5 solves this task by applying agglomerative clustering methods. We concentrate on developing proper representations of speaker segments for clustering, research different similarity measures for joining speaker segments and explore different stopping criteria for clustering that would result in a minimization of the overall diarization error of such systems.

We realize two baseline systems. The first is a standard approach using a bottom–up agglomerative clustering principle with the BIC as a merging criterion. In the second system the alternative approach is applied, also using bottom–up clustering, but the representations of the speaker segments and the merging criteria are different. In this approach the speaker segments are modeled by GMMs. In the clustering procedure during the merging process universal background models (UBMs) are transformed into speaker–segment GMMs using the MAP adaptation technique. This is the common approach for modeling speakers in the speaker–identification and verification tasks. The merging criterion in such clustering is a cross log–likelihood ratio (CLR). We explored other similarity measures and found that the modified BIC measure performed the best of all the tested measures.
In the next approach a fusion speaker clustering system is developed, where the speaker segments are modeled by acoustic and prosody representations. The idea here is to additionally model the speaker prosody characteristics and add it to basic acoustic information estimated from the speaker segments. We construct 10 basic prosody features derived from the energy of the audio signals, the estimated pitch contours, and the recognized basic speech units. In this way we impose higher–level information in the representations of the speaker segments, which leads to improved clustering of the segments in the case of similar speaker acoustic characteristics or poor acoustic conditions. By adding prosody information to the basic acoustic features the baseline clustering procedure has to be changed to work in the fusion of both representations.

In the second part of Chapter 5 we explore different stopping criteria for speaker clustering to find the final number of clusters that tend to minimize the overall diarization error. We propose two alternative criteria to the baseline criterion, which is usually defined by introducing the stopping threshold estimated from the evaluation data. Such thresholds should be set in advance and should match the acoustic conditions of the training and working environments. With our approaches we tried to overcome this. The first proposed criterion is based on an overall BIC measure and it works well together with the BIC as a merging criterion. The second approach is more suitable for the task of speaker clustering, since it tries to find the optimal number of clusters by inspecting the relative difference of the diarization error produced by two different clustering procedures.

We performed two groups of evaluation experiments where the diarization error rate (DER) was used as an assessment measure in all our experiments. In the first group an ideal segmentation of audio data was assumed and just speaker clustering was performed on manually annotated speaker segments, while in the second group of experiments speaker clustering was applied on automatically derived segments. In the first case the performances of the clustering procedures alone were studied, while in the latter case an assessment of all the speaker diarization tasks was carried out. Although the evaluation results varied among the different experiments, it could be concluded that speaker clustering and diarization systems, where the segments are modeled by speaker–oriented representations (speaker GMMs, prosody features), performed more stably and reliably than the baseline systems, where segments are represented just by acoustic information. The best overall results were achieved with the fusion system where clustering was performed by joining the acoustic and prosody features.

In conclusion, we summarize and discuss the presented methods and their results. We also provide an overview of the possible uses of the proposed methods in various speech applications and demonstrate the integration of the speaker–diarization procedures into an audio–indexing system. At the end, some directions for future research and improvements to the proposed methods are given.

Kazalo

Slike

Tabele

Seznam pogosto uporabljenih kratic

BIC	.....	Bayesov informacijski kriterij, kriterij BIC
CLR	.....	navzkrižni kriterij razmerij logaritmov verjetnostnih ocen (LLR)
CMVN	.....	normalizacija z izničevanjem skupnega povprečja in variance kepstralnih značilk
CVS	.....	glasovne enote: soglasnik (C), samoglasnik (V), premor (S)
DER	.....	mera napake med ujemanjem referenčnih ter samodejno pridobljenih in označenih segmentov, mera DER
$\mathbf{f_0}$	.....	višina osnovnega tona v govornem signalu
FW	.....	postopek prileganja značilk k normalnim porazdelitvam
GMM	.....	model kombinacije Gaussovih porazdelitev, GMM model
GNG	.....	govor/ne–govor
HMM	.....	prikrit Markovov model, HMM model
KL	.....	Kullback-Leiblerjeva (divergenčna) mera verjetnostne podobnosti
KL2	.....	simetrična Kullback-Leiblerjeva (divergenčna) mera verjetnostne podobnosti
LLH	.....	logaritem verjetnostne ocene
LLR	.....	razmerje logaritmov verjetnostnih ocen
LVCSRs	.....	sistem za razpoznavanje tekočega govora velikega števila različnih besed
MAP	.....	postopek prilagajanja (GMM) modelov z večanjem aposteriornih verjetnosti
MFCC	.....	koeficienti melodičnega kepstra, MFCC koeficienti
PRC	.....	natančnost pri ocenjevanju segmentacije
RCL	.....	priklic pri ocenjevanju segmentacije

SAG	.....	sprememba po govorcih in v akustičnem ozadju
SG	.....	pri segmentaciji sprememba po govorcih, pri rojenju segmenti po govorcih
SVM	.....	metoda podpornih vektorjev
VUS	.....	glasovne enote: zveneči glas (V), nezveneči glas (U), premor (S)
WER	.....	napaka razpoznavanja govora
UBM	.....	splošen (GMM) model govora
ZCR	.....	število prehodov signala skozi nič

1 Uvod

V uvodnem poglavju bomo predstavili temo doktorske disertacije in opredelili glavne cilje raziskovalnega dela, ki smo se jim posvetili v doktorski disertaciji. Podali bomo tudi splošen pregled področja obdelave in analize informativnih oddaj z uporabo govornih tehnologij, kjer se bomo osredotočili predvsem na pregled temeljnih del iz ožjega področja teme disertacije. Zaključili bomo s pregledom vsebine disertacije.

1.1 Tema disertacije

V okviru doktorske disertacije smo se osredotočili predvsem na obdelavo in analizo zvočnih posnetkov informativnih oddaj z uporabo postopkov govornih tehnologij. Glavni namen obdelave zvočnih posnetkov je bil, da bi z uporabo postopkov govornih tehnologij samodejno organizirali in označili zvočne posnetke informativnih oddaj tako, da bi bili primerni za pridobivanje, iskanje in združevanje različnih tipov informacij, ki so posredovane preko zvočnih zapisov informativnih oddaj. Na ta način bi osnovne zvočne posnetke opremili z dodatno informacijo, ki bi bila primerna za nadaljnjo obdelavo informativnih oddaj. Kljub temu, da je možno predvsem pri televizijskih informativnih oddajah pridobivati tudi druge tipe podatkov, npr. video posnetke, smo vse ostale podatke uporabljali le kot dopolnilno informacijo osnovnim zvočnim podatkom.

Osnovna naloga obdelave zvočnih posnetkov informativnih oddaj z uporabo govornih tehnologij je pretvorba govornih podatkov iz zvočne v tekstovno obliko. To dosežemo s postopki razpoznavanja govora. Vsi ostali postopki služijo bodisi za predpripravo zvočnih podatkov za razpoznavanj bodisi za pridobivanje dodatne informacije osnovnim tekstovnim prepisom pridobljenih iz razpoznavanja govora bodisi za nadaljnjo obdelavo tekstovne in dodane informacije za namene samodejnega strukturiranja in pridobivanja informacij iz informativnih oddaj. Zadnjo skupino postopkov lahko že uvrstimo med postopke jezikovnih tehnologij in se v disertaciji z njimi nismo ukvarjali. Posvetili smo se prvima dvema skupinama postopkov: predobdelavi zvočnih posnetkov za namene razpoznavanja govora in strukturiranju posnetkov za pridobivanje dodatne informacije. Na ta način smo želeli pripraviti zvočne posnetke informativnih oddaj za razpoznavanje govora, hkrati pa smo jih želeli opremiti z dodatno informacijo, da bi bili primerni za vključevanje v multimedijske arhive informativnih oddaj. Zato smo se v disertaciji ukvarjali predvsem s postopki organizacije zvočnih podatkov za indeksacijo zvočnih posnetkov *(ang. audio indexing)*, ki bi bili primerni za iskanje in pridobivanje vsebinskih informacij iz zvočnih zapisov informativnih oddaj *(ang. content–based audio retrieval)*.

Naloge, ki smo jih reševali v disertaciji, so bile definirane v okviru mednarodnega projekta *Rich Transcription* [Fiscus-05] z nadaljevanjem v projektu *CHIL* [Waibel-04] in v projektu *ESTER* [Istrate-05]. V projektu *Rich Transcription*, ki se je začel leta 2002, so želeli osnovne transkripcije govora informativnih oddaj opremiti z dodatnimi informacijami o govorcih in kvaliteti zvočnih posnetkov. V projektu *CHIL* so posnetkom informativnih oddaj dodali še zvočne in video posnetke poslovnih srečanj, zato so si zastavili organiziranje vsebin širše s pomočjo večmodalnih predstavitev podatkov. V projektu *ESTER* pa je bil cilj strukturiranje zvočnih podatkov za namene iskanja in sledenja govorcev v podatkovnih zbirkah zvočnih posnetkov.

V okviru doktorske disertacije smo se ukvarjali predvsem s samodejnim označevanjem in organiziranjem zvočnih podatkov na podlagi informacije o govorcih, s čimer smo želeli pripraviti vse potrebno za indeksacijo zvočnih posnetkov za namene iskanja in sledenja govorcev v multimedijskih podatkovnih zbirkah informativnih oddaj.

1.2 Cilji raziskovalnega dela

Osnovna naloga postopkov iz disertacije je bila priprava zvočnih posnetkov informativnih oddaj za razvoj sistema za razpoznavanje govora in za razvoj sistema za iskanje in pridobivanje vsebinskih informacij iz zvočnih posnetkov informativnih oddaj. Pri tem smo se usmerili predvsem k postopkom predobdelave zvočnih posnetkov, s katerimi bi zagotavljali bolj zanesljivo in učinkovito delovanje obeh sistemov.

Za razvoj takšnih sistemov je potrebno najprej pridobiti ustrezno označene podatkovne zbirke posnetkov informativnih oddaj. Zato smo si v prvi fazi zastavili cilj, da bi pridobili in ustrezno označili podatkovno zbirko zvočnih posnetkov informativnih oddaj v slovenskem jeziku, ki bi jo lahko uporabljali za razvoj postopkov iz različnih področij govornih tehnologij. Naš namen je bil predvsem izgradnja sistema za razpoznavanje tekočega govora velikega števila različnih besed (*ang. large vocabulary continuous speech recognition system, LVCSRs).* Zato je bilo potrebno pridobiti velike količine govornega materiala. Po drugi strani pa smo želeli pridobiti tudi ustrezne podatkovne zbirke, s katerimi bi bil mogoč razvoj postopkov za pridobivanje vsebinskih informacij, ki bi bile neodvisne od jezika. V ta namen smo želeli pridobiti večjezikovno podatkovno zbirko zvočnih posnetkov informativnih oddaj, ki bi bila pestra tako po vsebini kot tudi po akustičnih lastnostih zvočnih podatkov.

Osnovni cilj raziskovalnega dela v okviru doktorske disertacije je bil predvsem razvoj postopkov za strukturiranje zvočnih posnetkov, ki bi bili primerni za nadaljnjo obdelavo v postopkih pridobivanja različnih vsebinskih informacij iz informativnih oddaj. Tu smo se predvsem omejili na organiziranje zvočnih posnetkov glede na informacije o govorcih. Osnovne naloge, s katerimi smo se ukvarjali v disertaciji, lahko razdelimo na naslednja področja:

- segmentacija zvočnih posnetkov na govorne in ne–govorne dele,
- segmentacija zvočnih posnetkov glede na akustične spremembe ali zamenjave govorcev,
- razvrščanje segmentov po govorcih s postopki rojenja.

Osnovni cilj postopkov je bil, da bi organizirali zvočne posnetke tako, da bi združevali skupaj tiste govorne odseke zvočnih posnetkov informativnih oddaj, ki pripadajo istim govorcem. Na ta način bi osnovne zvočne zapise informativnih oddaj dopolnili z informacijo, primerno za iskanje in sledenje govorcev v posnetkih.

Problem pridobivanja in združevanja segmentov po govorcih je bil prvič definiran v okviru projekta *Rich Transcription* [Fiscus-05] v evaluacijah *"Who spoke when"*. Vključuje vse tri osnovne naloge, ki smo jih reševali v okviru doktorske disertacije. V zvočnih posnetkih je potrebno najprej poiskati in določiti govorne dele, nato izvesti segmentacijo govornih odsekov na homogene enote - segmente, ki jih potem združujemo v posamezne skupine, ki pripadajo samo enemu govorcu. Prvi del naloge rešujemo s postopki segmentacije zvočnih posnetkov na govorne in ne–govorne dele, nato pa pri segmentaciji govornih delov poskušamo razdeliti govorne odseke v takšne segmente, ki

bi bili primerni za združevanje po govorcih. V disertaciji smo se posvetili postopkom segmentacije po govorcih, torej razdelitvi govornih odsekov na segmente, ki pripadajo posameznim govorcem. Zadnji del naloge, ki zajema razvrščanje segmentov, pa običajno izvajamo s postopki rojenja z združevanjem prej pridobljenih segmentov. Tudi tu je osnova za združevanje informacija o govorcih.

Osnovne zahteve pri razvoju vseh treh skupin postopkov so bile, da bi bili postopki neodvisni od jezika, da bi jih lahko vključevali v različne sisteme nadaljnje obdelave posnetkov informativnih oddaj in da bi izboljšali delovanje postopkov v primeru različnih akustičnih pogojev. Vzporedno z razvojem postopkov smo želeli izboljšati tudi postopke vrednotenja posameznih nalog, s čimer smo želeli ustrezno oceniti in primerjati postopke v različnih pogojih delovanja.

1.3 Pregled področja teme disertacije

V nadaljevanju bomo podali širši pregled področij obdelave in analize informativnih oddaj z uporabo govornih tehnologij, kjer bomo ustrezno umestili raziskovalno delo doktorske disertacije. V zadnjem delu pa bomo podali bolj natančen pregled področja, s katerim smo se ukvarjali v disertaciji.

1.3.1 Splošen pregled obdelave in analize informativnih oddaj z uporabo govornih tehnologij

Raziskave obdelave in analize informativnih oddaj z uporabo govornih tehnologij so postale aktualne koncem devetdesetih let prejšnjega stoletja, ko se je predvsem za angleško govoreče področje začelo večje število projektov na temo zbiranja in obdelave informativnih oddaj.

Pod okriljem združenj LDC[1] ter sponzorstvom ameriške agencije za standarde in tehnologijo, NIST[2], in agencije DARPA[3] [Pallett-02] so v letih 1996-98 pripravili prvo večjo zbirko označenih informativnih oddaj *(ang. English Broadcast News Speech, HUB-4)*, ki obsega približno 200 ur označenih posnetkov različnih informativnih oddaj v angleškem jeziku [Graff-02]. Vzporedno s tem so pridobili še 30 ur zvočnih posnetkov v kitajskem in španskem jeziku (Hub-4-NE). Osnovno zbirko v angleškem jeziku so v okviru različnih nadaljevalnih projektov (Rich Transcription, [Fiscus-05] CHIL, [Waibel-04]) razširili na nekaj tisoč ur posnetkov, kjer so poleg posnetkov informativnih oddaj dodali še posnetke poslovnih sestankov, predavanj ipd. Sočasno z razvojem tehnologij obdelave takšnih podatkov za angleški jezik so nastajale podatkovne zbirke informativnih oddaj tudi v drugih jezikih: nemškem [Macherey-02, BN-DWK-99], francoskem [Galliano-05], japonskem [Furui-98], italijanskem [Federico-00], portugalskem [Meinedo-01] in drugih.

[1] Lingustic Data Consortium, http://www.ldc.upenn.edu/
[2] National Institute of Standards and Technology, http://www.nist.gov/
[3] Defense Advanced Research Projects Agency, http://www.darpa.mil/

Prvotni namen takšnih podatkovnih zbirk je bil preučevanje in prenos obstoječih postopkov obdelave in razpoznavanje govora v kontroliranem okolju na večje sisteme razpoznavanja kompleksnih govornih podatkov z namenom samodejnega označevanja informativnih oddaj *(ang. automatic broadcast news transcription)* [Woodland-02, Beyerlein-02, Chen-02, Gauvain-02]. Takšni sistemi so predstavljali osnovo za razvoj tehnologij za samodejno pridobivanje informacij neposredno iz zvočnih podatkov [Makhoul-00, Federico-00, Meinedo-03a] ali v kombinaciji z drugimi viri, npr. videom [Kemp-98, Viswanathan-00]. V zadnjem času se je raziskovanje na tem področju usmerilo predvsem v razvoj sistemov za samodejno indeksacijo zvočnih posnetkov *(ang. automatic audio indexing)* [Makhoul-00, Olive-00, Neti-00] in detekcijo vsebin informativnih oddaj *(ang. topic detection)* [Walls-99, Wayne-00, Leek-00].

V primerjavi s klasičnimi nalogami, ki jih rešujemo z uporabo govornih tehnologij, predstavlja samodejno označevanje in obdelava zvočnih posnetkov informativnih oddaj veliko zahtevnejši problem. Zvočni posnetki pridobljeni v nekontroliranih pogojih vključujejo večplastne akustične in jezikovne informacije. Zvočni posnetki so v povprečju bistveno daljši, saj trajanja posameznih oddaj lahko dosežejo tudi čas ene ure ali celo več, akustične vsebine pa so tipično zelo pestre tako po vsebinski plati, številu govorcev, načinu govora, akustičnem ozadju in delih akustičnih vsebin, ki niso govor. Po drugi strani pa je tudi informativna vsebina takega tipa podatkov veliko obsežnejša in poleg informacije, ki bi jo podal tekstovni prepis govora, obsega še podatke o govorcu, uporabljenem jeziku in načinu govora, tipu vsebine, aktualnem času itd. Pri televizijskih oddajah je možno dodatne podatke pridobiti še iz slikovnega gradiva in jih uskladiti z zvočnim delom. V tem primeru govorimo o večmodalni obdelavi. Zato s klasičnimi postopki obdelave in razpoznavanja govornih signalov pridobljenih v kontroliranih pogojih ne moremo zagotoviti natančne analize posnetkov informativnih oddaj [Woodland-02].

Za razvoj in preizkušanje različnih pristopov pri večmodalnih obdelavah govorjenega jezika, ki so namenjeni za pridobivanje informacij iz multimedijskih virov, je potrebno pridobiti ustrezne zbirke večmodalnih podatkov. Pridobivanje predvsem pa označevanje takšnih podatkov je zahteven in dolgotrajen proces. Označevanje zvočnih posnetkov informativnih oddaj [LDC-00] mora vključevati tako transkripcijo govornih delov, kot tudi oznake govorca, ozadja, kvalitete posnetka in tipa govora, posebne oznake za negovorne dele, tuj jezik, oznake za tip in vsebino novic ipd. V ta namen so bila razvita posebna orodja za označevanje takšnega tipa podatkov; med njimi je najbolj razširjeno in uporabljeno orodje *Transcriber* [Barras-01].

Raziskave in uporaba sistemov za samodejno obdelavo takega tipa podatkov so zelo različne in jih lahko razdelimo na naslednja področja:

- samodejna transkripcija informativnih oddaj,
- indeksacija zvočnih posnetkov,
- sinhronizacija večmodalnih podatkov, predvsem teksta z avdio/video posnetki,
- podnaslavljanje (delov) informativnih oddaj,
- detekcija in sledenje posameznim vsebinam ali novicam informativnih oddaj,

- pridobivanje in iskanje informacij iz multimedijskih virov ter
- identifikacija in obdelava večjezikovnih vsebin informativnih oddaj.

V nadaljevanju bomo opisali osnovne probleme, s katerimi se ukvarjamo pri naštetih področjih obdelave in analize informativnih oddaj, podali bomo pregled temeljnih del ter ustrezno v ta področja umestili postopke, s katerimi smo se ukvarjali v doktorski disertaciji.

1.3.1.1 Samodejno pridobivanje transkripcij informativnih oddaj

Tu gre v bistvu za sisteme razpoznavanja govora, uporabljene na segmentiranih in drugače ustrezno obdelanih govornih posnetkih. Sistemi za razpoznavanje govora so ključnega pomena za razvoj nadaljnjih sistemov za samodejno pridobivanje informacij iz informativnih oddaj. Zaradi narave posnetkov informativnih oddaj, ki lahko vključujejo različne oblike in tipe akustičnih dogodkov, različno kvalitetne posnetke, veliko število govorcev, različne načine govora, širok spekter različnih vsebin in s tem bogato jezikovno informacijo, so takšni sistemi kompleksni in združujejo najnovejša dognanja in tehnologije govorjenega jezika.

Za izvedbo sistemov samodejne transkripcije poleg natančno označenih govornih posnetkov potrebujemo tudi velike korpuse besedil za izgradnjo ustreznih jezikovnih modelov, ki praviloma vključujejo nekaj milijonov besed. Takšni sistemi so odvisni od jezika razpoznavanja. To pomeni, da je potrebno pridobiti ustrezne zbirke govorjenih informativnih oddaj ter velike korpuse besedil v jezikih, za katere gradimo takšne sisteme. Zato smo tudi v okviru raziskovalnega dela doktorske disertacije začeli s pridobivanjem in označevanjem slovenske zbirke zvočnih posnetkov informativnih oddaj in z zbiranjem tekstovnih prepisov informativnih oddaj.

Sistemi razpoznavanja govora so v glavnem zasnovani na osnovi prikritih Markovovih modelov (HMM modeli) z izboljšanimi metodami prilagajanja modelov in tehnikami iskanja, prilagojenimi velikim besediščem razpoznavanja in velikim jezikovnim modelom [Chen-02, Gauvain-02, Woodland-02, Beyerlein-02]. Ravno kompleksnost zvočnih podatkov informativnih oddaj zahteva predobdelavo takšnih posnetkov s postopki, s katerimi smo se ukvarjali v okviru doktorske disertacije. Tako se je izkazalo, da s postopki združevanja segmentov po govorcih, ki sledijo samodejni segmentaciji zvočnih posnetkov, dosežemo bistvene izboljšave razpoznavanja govora. V tem primeru se namreč v fazi prilagajanja modelov splošne modele govora nadomesti z modeli prilagojenimi na posameznega govorca *(ang. speaker–adapted training, SAT)*. Na ta način se bistveno izboljšajo rezultati razpoznavanja [Kubala-97, Siegler-97, Zhang-02].

Druga pomembna lastnost sistemov za razpoznavanje je, da se ločijo na sisteme, ki delujejo v realnem času in takšne, kjer je čas razpoznavanja večkratnik časa trajanja posnetkov. Napaka razpoznavanja *(ang. word error rate, WER)* za slednje sisteme je v povprečju med 10 in 25%, [Pallett-99]. Ti sistemi so v glavnem razviti za angleški jezik [Chen-02, Beyerlein-02, Gauvain-02, Woodland-02], v zadnjem času pa se intenzivno razvijajo tudi za druge, predvsem evropske jezike: nemščino, francoščino in nizozemščino v okviru projekta Olive [Olive-00] in njegovem nadaljevanju v projektu

MUMIS [Saggion-04], v italijanskem [Federico-00] in portugalskem [Meinedo-03a] jeziku. Poseben izziv predstavljajo sistemi, ki delujejo v realnem času in praviloma dosegajo slabše rezultate. Namenjeni so v glavnem sprotnemu podnaslavljanju informativnih oddaj. Pri takšnih sistemih je potrebno prilagoditi vse postopke obdelave signalov, segmentacije, razvrščanja in razpoznavanja sprotnemu delovanju. Zato je potrebno dodatno optimizirati predvsem postopke razpoznavanja, tj. akustične modele ter izračune in postopke iskanja optimalnih poti skozi grafe akustičnih in jezikovnih modelov. Tak je npr. sistem [Saraclar-02], kjer so dosegli na bazi HUB-4 samo 22% WER.

1.3.1.2 Samodejna indeksacija zvočnih posnetkov

Pri indeksaciji zvočnih posnetkov gre za avtomatično označevanje zvočnih posnetkov glede na določeno tematsko področje, osebo ali govorca, časovni okvir, jezik ipd. z namenom izgradnje podatkovne zbirke, namenjene pridobivanju in iskanju informacij iz zvočnih virov.

Glede na tip informacije, ki jo iščemo, se takšni sistemi delijo na več skupin:

- Indeksacija glede na vrsto jezika, kjer gre v bistvu za problem identifikacije jezika. Tu se uporabljajo postopki identifikacije izvedeni z uporabo jezikovno odvisnih razpoznavalnikov, ki temeljijo na monofonskih akustičnih in n-gramskih jezikovnih modelih [Zissman-96], kar pomeni, da za učenje takšnih modelov potrebujemo zbirke ustrezno označenih akustičnih posnetkov jezikov, ki jih razpoznavamo.

- Indeksacija po govorcih, kjer organiziramo zbirko zvočnih in/ali video posnetkov glede na govorce, ki so prisotni v podatkih. V prvem koraku izvedemo segmentacijo vsakega posnetka po govorcih. V drugem koraku se izvaja združevanje segmentov posameznih govorcev znotraj avdio/video dokumenta v enotno listo govorcev *(ang. speaker segments tying)* [Meignier-02]. V zadnjem koraku pa se izdela indeks govorcev *(ang. speaker-based index)* za učinkovito iskanje govorcev v zbirki.

- Sledenje posameznemu govorcu; tu iščemo segmente - dele posnetkov v avdio/video zbirki, kjer govori iskani govorec. Postopki sledenja so tu v glavnem izvedeni s statističnimi modeli, običajno se uporablja en model za iskanega govorca *(ang. target speaker model)* in en ali več modelov za ozadja *(ang. background model)*, s katerim zajamemo akustične lastnosti vseh ostalih govorcev [Magrin-99, Bonastre-00].

- Indeksacija po ključnih besedah; tu gre po zgledu medmrežnih iskalnikov za detekcijo ključnih besed *(ang. keyword spotting)* iz zvočnih posnetkov. V glavnem se uporabljata dva pristopa za iskanje ključnih besed iz zvočnih posnetkov, in sicer klasični z uporabo LVCSRs sistemov za razpoznavanje, s katerimi iz akustičnega signala pridobimo najbolj verjetno zaporedje besed in med njimi iščemo ključne besede [Weintraub-93]. Alternativni pristop pa je opisan v [Manos-97]. Tu se uporablja samo akustične modele ključnih besed in skupne modele ostalih besed ob določeni ključni besedi *(ang. filler models)*. V prvem primeru potrebujemo ogromno označenega materiala za učenje parametrov razpoznavalnika,

medtem ko v drugem precej manj. Rezultati detekcije ključnih besed so v prvem primeru boljši [Manos-97].

- Indeksacija po vsebinah novic informativnih oddaj, kjer se uporabljajo tehnike detekcije in sledenja posameznim vsebinam, ki jih bomo predstavili v nadaljevanju.

Tudi pri indeksaciji in sledenju posameznim govorcem v zvočnih posnetkih uporabljamo postopke, s katerimi smo se ukvarjali v disertaciji. V obeh primerih je potrebno zvočne posnetke ustrezno pripraviti za nadaljnjo obdelavo. Zato se tu uporabljajo tako postopki segmentacije zvočnih posnetkov na govorne in ne–govorne dele, kot tudi segmentacija govornih delov po govorcih in razvrščanje segmentov s postopki rojenja. Učinkovito in zanesljivo delovanje teh postopkov je bistvenega pomena za uspešno delovanje sistemov indeksacije zvočnih posnetkov.

Samodejna indeksacija predstavlja osnovo za izgradnjo sistemov za iskanje informacij po multimedijskih arhivih [Wactlar-99] in za pridobivanje informacij iz multimedijskih vsebin [Makhoul-00, Kemp-98, Gauvain-03].

1.3.1.3 Sinhronizacija večmodalnih podatkov

V primeru sinhronizacije informacij pridobljenih iz različnih virov se v okviru informativnih oddaj osredotočamo predvsem na sinhronizacijo besedila z avdio in/ali video posnetki. V tem primeru izvajamo časovno poravnavo avdio/video signala s tekstovno predlogo *(ang. text and audio/video data alignment)*, ki je podana vnaprej, vendar ne predstavlja nujno natančnega zapisa govora v zvočnem posnetku. Tako npr. za arhive zvočnih posnetkov, pa tudi za nekatere oddaje, ki sicer potekajo v živo, vendar po vnaprej pripravljeni besedilni predlogi (bran govor), pogosto že obstajajo besedilne predloge, ki podajajo bolj ali manj pravilen besedni prepis govora v tekstovni obliki. Ena izmed možnosti pridobivanja takih prepisov je uporaba teleteksta. V takih primerih lahko s postopki vsiljenega prileganja *(ang. forced alignment)* [Moreno-98], ki uporabljajo akustične modele osnovnih govornih enot izbranega jezika, določene s HMM modeli, leksikon s fonetičnimi prepisi besed, zvočni posnetek in njegov tekstovni prepis, dosežemo časovno poravnavo delov besedila z govornimi segmenti. Ovire pri takih postopkih predstavljajo nenatančni tekstovni prepisi in elementi spontanega govora, kot so npr. glasni vdihi in izdihi, smeh, premori, napačna izgovorjava, zatikanja pri izgovorjavi in ponavljanja. Za zagotovitev zanesljivega delovanja postopkov sinhronizacije je potrebno take dogodke predvideti in jih ustrezno akustično modelirati. Bistvenega pomena je tu detekcija govornih in ne–govornih delov v zvočnih posnetkih, saj nam uspešna lokalizacija ne–govornih delov zagotavlja uspešno poravnavo govornih delov. Zato je potrebno v takšnih sistemih predhodno segmentirati zvočne posnetke na govor in ne–govor.

Postopki sinhronizacije podatkov so izpeljani neposredno iz sistemov za razpoznavanje govora, vendar so v kombinaciji s segmentacijo in bimodalnimi tehnikami prilagojeni poravnavi večurnih avdio in/ali video posnetkov. Uporabljajo se zlasti v sistemih za samodejno podajanje vsebin, pridobljenih iz različnih multimedijskih vi-

rov [Makhoul-00, Maybury-99, Wactlar-99, Gauvain-00]. Primer postopkov poravnave teksta in zvočnih podatkov v povezavi z indeksacijo pa je podan v [Biatov-03].

1.3.1.4 Podnaslavljanje informativnih oddaj

Sistemi samodejnega podnaslavljanja informativnih oddaj vključujejo postopke segmentacije zvočnih posnetkov, razpoznavanja govora in sinhronizacije večmodalnih tokov podatkov. Bistvena zahteva takšnih sistemov je, da tečejo v realnem času, kar znatno poveča težavnost problema. Prvi tak sistem so razvili v Bellovih laboratorijih za japonski jezik [Siohan-01], kjer gre za podnaslavljanje različnih tipov oddaj znotraj ene vrste informativne oddaje. Z razvojnim sistemom so dosegli odlične rezultate razpoznavanja govora pri voditeljih informativnih oddaj (preko 90%) in rezultate med 78% in 90% za ostale tipe govora [Siohan-01]. Podoben sistem so zgradili tudi v AT&T-jevih raziskovalnih laboratorijih, kjer so dosegli rezultate razpoznavanja okoli 80% [Saraclar-02]. Manjši sistemi so razviti v glavnem za samodejno podnaslavljanje določenih tipov oddaj zaključenih vsebin npr. športnih oddaj, vremenskih napovedi [Žibert-00], finančnih novic, ipd.

Namen uporabe takšnih sistemov je predvsem pomoč gluhim in naglušnim osebam za nemoteno spremljanje dnevno–informativnih oddaj in oddaj v živo ter avtomatizacija obstoječih sistemov za podnaslavljanje preko teleteksta, kjer so se pojavili že prvi specializirani komercialni produkti [WinCAPS-06, Aurix-03].

1.3.1.5 Detekcija in iskanje zaključenih vsebin novic

Detekcija zaključenih vsebin novic *(ang. topic detection)* pomeni, da v segmentiranem podatkovnem viru odkrijemo in med sabo povežemo dele (segmente) z istega ali sorodnih vsebinskih področij. Intenzivno raziskovanje na tem področju se je začelo v letih 1998/99, ko se je pod okriljem DARPA[4] začelo zbiranje, označevanje, razvoj in vrednotenje postopkov detekcije in sledenja vsebinam novic (ang. *topic detection and tracking*), pridobljenih iz različnih virov in v različnih jezikih [Wayne-00]. V splošnem lahko probleme detekcije in sledenja novic, pridobljenih iz informativnih oddaj, razdelimo na štiri raziskovalna področja:

- Segmentacija na področja zaključenih vsebin novic *(ang. story segmentation)*. Pri tej nalogi moramo s postopki v zveznem vhodnem besedilnem toku odkriti meje med posameznimi novicami. Predpogoj segmentacije po novicah zvočnih posnetkov informativnih oddaj je predhodna segmentacija zvočnih posnetkov [Shriberg-00].

- Sledenje zgodbe novice *(ang. topic tracking)*. Tu na podlagi že prepoznanih dogodkov, ki govorijo o določeni temi, sledimo zgodbi v besedilu, ki ga trenutno obdelujemo. Postopki sledenja vsebin so v glavnem izvedeni s statističnimi modeli [Yamron-00] in s postopki razvrščanja [Leek-00].

[4]Glej opombo 3 na strani 4.

- Detekcija vsebin novic *(ang. topic detection)*. To področje vključuje sprotno ali retrospektivno razpoznavanje vsebin v toku besedila, postopke samodejne izdelave povzetkov *(ang. automatic summarization)* in postopke organiziranja vsebin za nadaljnjo uporabo v sistemih za podajanje informacij. Pregled raziskovalnih dosežkov na področju obdelave in pridobivanja informacij dnevno–informativnih oddaj je v [Walls-99], samodejna izdelava povzetkov vsebin iz zvočnih podatkov je opisana v [Hori-02], organizacija zvočnih posnetkov za nadaljnjo uporabo v informacijskih sistemih pa v [Gauvain-03].
- Povezovanje področij sorodnih vsebin novic *(ang. link detection)*. V tem primeru s pomočjo postopkov segmentacije in detekcije vsebin razvrščamo in združujemo novice v večje razrede v glavnem s postopki rojenja tekstovne informacije [Brown-99].

Naloge detekcije in iskanja zaključenih vsebin novic vključujejo v glavnem postopke obdelave jezikovne informacije, ki spadajo v širše področje jezikovnih tehnologij, vendar so osnova vsem postopkom tekstovni podatki, ki jih pridobimo z razpoznavanjem govora iz zvočnih posnetkov informativnih oddaj. Za uspešno delovanje takšnih sistemov je zato potrebno zagotoviti ustrezno segmentacijo razpoznanega govora na smiselne enote (običajno so to stavki), ki jih lahko združujemo in povezujemo v zaključene vsebine novic.

1.3.1.6 Obdelava večjezikovnih vsebin informativnih oddaj

Združevanje večjezikovnih multimedijskih zbirk predstavlja nov izziv v obdelavi govorjenih informativnih oddaj. Takšne zbirke so v glavnem uporabne za preizkušanje robustnosti od jezika neodvisnih postopkov obdelave in analize zvočnih posnetkov. Takšna je npr. zbirka informativnih oddaj v devetih evropskih jezikih, ki je nastajala v okviru projektne skupine COST278[5], [Vandecatseye-04], in bo opisana v naslednjem poglavju. Primerna je za preizkušanje postopkov segmentacije in indeksacije zvočnih posnetkov, detekcije vsebin in identifikacije jezika.

Na inštitutu LIMSI v Franciji so zgradili večjezikovni sistem za pridobivanje informacij iz informativnih oddaj [Lamel-02], ki pokriva pet glavnih evropskih jezikov ter kitajski in arabski jezik. Sistem je namenjen raziskovanju na področju večjezikovne indeksacije zvočnih posnetkov *(ang. multi-lingual audio indexing)* in preučevanju prenosljivosti postopkov razpoznavanja govora med posameznimi jeziki. V okviru projekta MUSA [MUSA-02] pa so razvili sistem, ki je namenjen samodejnemu podnaslavljanju in prevajanju multimedijskih vsebin v treh jezikih: angleškem, francoskem in grškem.

1.3.1.7 Pridobivanje informacij iz multimedijskih vsebin

Rezultat združevanja in povezovanja postopkov obdelave in analize informacijskih oddaj z uporabo govornih tehnologij, večmodalnih tehnik in splošnih postopkov pridobivanja informacij iz podatkovnih zbirk predstavljajo sistemi pridobivanja informacij

[5]EU projekt COST Action 278: Spoken Language Interaction in Telecommunication

iz multimedijskih vsebin. Takšni sistemi so kompleksni in v večini primerov predstavljajo vrh tehnološkega napredka posameznih raziskovalnih skupin s področja obdelave in razpoznavanja govora, jezikovnega modeliranja in razumevanja ter pridobivanja znanja iz informacijskih vsebin. Takšni so npr. sistemi večjih raziskovalnih skupin [Johnson-01, Wactlar-99, Maybury-99], kjer gre za specializirane sisteme pridobivanja in obdelave informacij neposredno iz informativnih oddaj, ali pa širše zastavljeni sistemi pridobivanja informacij poljubnih multimedijskih vsebin v enem jeziku [Gauvain-00, Viswanathan-00] ali večjezikovnih vsebin [Saggion-04]. V zadnjem času se pojavljajo tudi že prvi komercialni sistemi, ki ponujajo samodejno indeksacijo in iskanje po različnih vsebinah multimedijskih posnetkov [Makhoul-00].

Osnova vsem tem sistemom je segmentacija in povezovanje segmentov multimedijskih posnetkov po različnih kriterijih združevanja, ki temeljijo tudi na postopkih, s katerimi smo se ukvarjali v doktorski disertaciji in bodo predstavljeni v nadaljevanju.

1.3.2 Pregled ožjega področja teme disertacije

Kot smo že nakazali pri splošnem pregledu obdelave in analize zvočnih posnetkov informativnih oddaj, lahko postopke, ki smo jih razvijali v doktorski disertaciji, uporabljamo v različnih sistemih govornih tehnologij. Praviloma jih izvajamo v začetnih fazah obdelave zvočnih posnetkov in služijo strukturiranju posnetkov na manjše, glede na področja uporabe, smiselne dele, ki so bolj primerni za nadaljnjo obdelavo.

Tako se pri detekciji govora ukvarjamo z razdelitvijo zvočnih posnetkov na govorne in ne–govorne dele, kar je predvsem uporabno v sistemih za razpoznavanje govora in govorcev. Tu je seveda potrebno izvajati razpoznavanje samo na govornih delih in ustrezna detekcija govora je bistvenega pomena za učinkovito in zanesljivo delovanje takšnih sistemov. Podobno je tudi pri segmentaciji zvočnih posnetkov po govorcih in/ali glede na akustične spremembe. Na ta način razdelimo daljše zvočne posnetke na odseke, ki pripadajo samo enemu govorcu z enakim akustičnim ozadjem. Združevanje segmentov tako razdeljenih posnetkov predstavlja osnovo za sisteme indeksacije in iskanja zaključenih vsebin v multimedijskih zbirkah podatkov. Učinkovito in zanesljivo delovanje postopkov segmentacije in združevanja segmentov tako zagotavlja pravilno indeksacijo zvočnih podatkov, kar omogoča boljšo organizacijo podatkov v zbirke, s čimer izboljšamo iskanje in sledenje določenim vsebinam v multimedijskih zbirkah. Druga pomembna lastnost takšnega strukturiranja podatkov pa je, da z uporabo teh postopkov pridobimo dodatno informacijo o vsebini in lastnostih multimedijskih (zvočnih) posnetkov.

V nadaljevanju bomo podali pregled dosedanjega dela na področjih, s katerimi smo se ukvarjali v disertaciji.

1.3.2.1 Samodejna detekcija govora v zvočnih posnetkih

Pri detekciji govora v zvočnih posnetkih rešujemo dva problema. Prvi zajema razvrščanje zvočnih odsekov (ali delov zvočnih posnetkov) na govor ali ne–govor, drugi

pa segmentacijo zvočnih posnetkov na podlagi teh lastnosti. Običajno se pri razvrščanju ukvarjamo z dvema nalogama: z iskanjem primernih predstavitev vzorcev za razvrščanje in s samimi postopki razvrščanja. V nadaljevanju si bomo pogledali, kakšne predstavitve zvočnih signalov se uporabljajo za detekcijo govora, katere postopke uporabljamo za razvrščanje in kako vse skupaj vgradimo v sisteme segmentacije.

Problem detekcije govora v zvočnih posnetkih lahko predstavimo kot segmentacijo zvočnih posnetkov na govorne in ne–govorne odseke, pri tem pa nas zanimajo samo govorni deli. Zato je bilo veliko študij narejenih predvsem za iskanje ustreznih predstavitev zvočnih signalov, ki bi bile primerne za detekcijo govora, razvrščanje pa je bilo omejeno na detekcijo govora glede na samo en tip ne–govornih posnetkov, predvsem glasbe. Zato se bomo tu omejili na pregled predstavitev, ki se uporabljajo za razvrščanje posnetkov na govor in glasbo, vendar lahko enake predstavitve uporabljamo tudi za razvrščanje govora glede na ostale ne–govorne pojave. Pri tem moramo poudariti, da so bile predstavitve zvočnih signalov načrtovane v glavnem za namene razvrščanja in ne za namene segmentacije zvočnih posnetkov.

Tako lahko ločimo dve skupini predstavitev ali značilk zvočnih signalov, s katerimi modeliramo govorne in ne–govorne pojave. Prva skupina predstavitev temelji na akustičnih lastnostih signalov, druga pa na specifičnih lastnostih govornih signalov v primerjavi z ne–govornimi (lahko so tudi akustične). Medtem ko so predstavitve iz prve skupine bolj splošne, so predstavitve iz druge skupine namenjene predvsem detekciji govora. Razvoj značilk za ločevanje med govorom in ne–govorom se je razvijal vzporedno z razvojem postopkov pridobivanja značilk za razpoznavanje govora. Tako je Greenberg [Greenberg-95] prvi predstavil značilke za ločevanje govora in glasbe, ki so temeljile na štetju prehodov signala skozi 0 *(ang. zero-crossing rate, ZCR)*. Te značilke je uporabil Saunders [Saunders-96] za predstavitve zvočnih signalov radijskih oddaj v sistemu za samodejno iskanje radijskih postaj na podlagi glasbe in govora. Podobne značilke, ki so temeljile na analizi signalov v času, je predstavil tudi Samouelian s sod. [Samouelian-98], ki pa jim je že dodal dve značilki iz frekvenčne predstavitve signalov. Prva, ki sta začela sistematično pridobivati značilke, primerne za modeliranje govora v sistemih za razvrščanje govora in glasbe, sta bila Scheirer in Slaney [Scheirer-97]. Obravnavala sta takšne značilke, ki se različno obnašajo v primeru govora in ostalih ne–govornih pojavov. Tako sta predlagala osnovne značilke, ki so temeljile na iskanju spektralnih središč *(ang. spectral centroid)*, merjenju spektralnega toka *(ang. spectral flux)*, štetju prehodov skozi 0 (ZCR), merjenju energije pri modulacijski frekvenci 4 Hz *(ang. 4 Hz modulation energy)*, ki ustreza frekvenci spreminjanja zlogov pri govoru, in merjenju deležev energije v nizkofrekvenčnih pasovih *(ang. percentage of low–energy frames)* časovno–frekvenčnih predstavitev signalov. V zadnjem času se v sistemih za ločevanje govora, glasbe in drugih ne–govornih pojavov najbolj pogosto uporabljajo predstavitve zvočnih signalov, ki jih uporabljamo tudi za razpoznavanje govora in govorcev. To so kepstralne značilke, ki jih pridobivamo na podlagi kratkočasovne frekvenčne analize signalov. Med njimi so najbolj razširjene značilke koeficientov melodičnega kepstra *(ang. mel–frequency cepstral coefficients, MFCC)* [Picone-93] in značilke koeficientov linearne predikcije *(ang. (perceptual) linear prediction coefficients, (P)LPC)* [Hermansky-90]. Uporaba teh predstavitev za ločevanje govora in ne–govora (predvsem glasbe) se je izkazala za učinkovito v kombinaciji z uporabo modelov kombinacije Gaussovih porazdelitev *(ang. Gaussian mixture models, GMM)*, [Logan-00] oziroma v sistemih, ki so

temeljili na HMM modelih, [Hain-98, Gauvain-02, Beyerlein-02, Ajmera-04]. Osnovni razlog za uporabo kepstralnih predstavitev signalov za detekcijo govora je tudi v tem, da so postopki detekcije govora običajno vključeni v večje sisteme za razpoznavanje govora, kjer poteka razpoznavanje skoraj izključno na podlagi kepstralnih značilk govora.

Druga skupina predstavitev zvočnih posnetkov za ločevanje govora in ne–govora temelji na drugačnih predpostavkah. Tu se problem ločevanja prevede na problem razvrščanja posnetkov v dva razreda, na razred govora in razred ne–govora. V tem primeru nas zato zanimajo takšne predstavitve, s katerimi lahko predstavimo zvočne signale tako, da jih modeliramo samo z dvema razredoma. Prvi poskus v tej smeri je izvedel Greenberg [Greenberg-95]. Najbolj uspešna pa sta bila Williams in Ellis [Williams-99], ki sta izpeljala značilke na podlagi opazovanja delovanja osnovnih sistemov za razpoznavanje govora. Tudi mi smo v svojem raziskovalnem delu sledili temu načinu izpeljave predstavitev zvočnih posnetkov, zato bomo natančnejši pregled in osnovne ideje tega tipa modeliranja zvočnih signalov predstavili v nadaljevanju, v tretjem poglavju.

V okviru raziskovalnega dela disertacije se nismo ukvarjali samo s predstavitvami zvočnih signalov, primernih za ločevanje govora in ne–govora, ampak tudi s postopki segmentacije zvočnih posnetkov na podlagi teh predstavitev. Večina omenjenih predstavitev je bila namreč preizkušana samo za razvrščanje že segmentiranih zvočnih posnetkov na govor in ne–govor. V primeru obdelave informativnih oddaj pa je potrebno dolge zvočne posnetke najprej razdeliti na dele, ki pripadajo govoru in ne–govoru, zato smo želeli razviti takšne predstavitve zvočnih signalov, ki bi bile primerne tudi za samo segmentacijo zvočnih posnetkov.

Dosedanje raziskovalno delo na področju segmentacije zvočnih posnetkov na govorne in ne–govorne dele je bilo usmerjeno predvsem k obravnavi in izvedbi postopkov segmentacije v okviru celotnih sistemov za razpoznavanje govora [Siegler-97, Woodland-02, Gauvain-02, Beyerlein-02] ali pa sistemov za sledenje in iskanje govorcev v zvočnih posnetkih [Zhu-05, Sinha-05, Žibert-05, Istrate-05, Moraru-05]. V večini primerov so se pri segmentaciji na govor in ne–govor uporabljale MFCC značilke za predstavitve zvočnih signalov in GMM ali HMM modeli za razvrščanje in segmentacijo zvočnih posnetkov. Alternativen pristop k segmentaciji je bil predlagan v [Lu-02], kjer so razvrščanje zvočnih posnetkov na govor in glasbo izvajali s pomočjo metode podpornih vektorjev *(ang. support vector machine, SVM)*.

Pri izgradnji samostojnega sistema segmentacije in razvrščanja zvočnih posnetkov, ki smo ga v okviru doktorske disertacije uporabljali kot referenčni sistem za detekcijo govornih delov, smo se zgledovali po [Ajmera-04]. Tu sta potekala segmentacija in razvrščanje segmentov istočasno. To smo dosegli z vključevanjem GMM modelov v mrežo HMM modelov, s katerimi smo v procesu segmentacije s postopkom Viterbijevega dekodiranja [Rabiner-89] dosegli razdelitev in označitev zvočnih posnetkov na govorne in ne–govorne dele. Razvili smo tudi alternativen pristop, kjer sta potekali segmentacija in razvrščanje segmentov ločeno. Tu smo najprej izvajali segmentacijo posnetkov glede na akustične spremembe v signalih, s katero se bomo podrobneje ukvarjali v poglavju 4, nato pa smo razvrščali segmente na govor in ne–govor s pomočjo GMM modelov.

Izvedba samostojnih sistemov segmentacije nam je tako omogočila razvoj in primerjavo

različnih predstavitev zvočnih signalov, primernih za modeliranje govora in ne–govora.

1.3.2.2 Samodejna segmentacija zvočnih posnetkov

Postopki segmentacije zvočnih posnetkov se ločijo glede na namen uporabe in glede na metode, uporabljene pri sami segmentaciji. Pri obdelavi zvočnih posnetkov informativnih oddaj ločimo postopke samodejne segmentacije glede na naslednja področja uporabe:

- segmentacija po govorcih in/ali glede na spremembe akustičnega ozadja; primerna je za nadaljnjo indeksacijo zvočnih posnetkov;
- segmentacija po stavkih na podlagi prozodične informacije [Shriberg-00] je osnova postopkom iskanja, detekcije in sledenja vsebinam informativnih oddaj;
- segmentacija na govor/šum/glasbo, pri govoru pa še na spol govorca ali samo na govor in ne–govor; predstavlja običajno prvi korak pri obdelavi zvočnega signala v sistemih za samodejne transkripcije in podnaslavljanje informativnih oddaj.

V disertaciji smo se ukvarjali s segmentacijo zvočnih posnetkov po govorcih in s segmentacijo na govorne in ne–govorne odseke, ki smo jo predstavili že v prejšnjem razdelku. S segmentacijo posnetkov po stavkih se nismo posebej ukvarjali, čeprav smo prozodično informacijo uporabljali pri razvrščanju segmentov po govorcih, kar bomo predstavili v nadaljevanju.

Pri samodejni segmentaciji zvočnih posnetkov po govorcih in/ali glede na spremembe akustičnega ozadja je osnovna naloga poiskati časovne meje v zvočnih signalih, s katerimi razdelimo zvočne posnetke na segmente glede na zamenjave govorcev *(ang. speaker change detection)* in/ali glede na spremembe v akustičnem ozadju *(ang. background change detection)*. Glede na metode, uporabljene pri takšni segmentaciji, razdelimo postopke segmentacije na dve skupini:

- *metode segmentacije s predhodnim učenjem*:
  Tu se v glavnem uporabljajo GMM modeli, v kombinaciji s HMM modeli, ki jih pridobimo na podlagi učnega materiala, segmentacija pa poteka s prileganjem modelov na predstavitve zvočnih posnetkov s postopki dinamičnega programiranja. Takšni so npr. sistemi opisani v [Kemp-00, Gauvain-02, Woodland-02]. V zadnjem času pa se je uveljavila tudi segmentacija z metodo podpornih vektorjev SVM, [Guo-03, Lu-02].
- *metode segmentacije s sprotnim odločanjem*:
  Tu se na podlagi podobnosti ali različnosti dveh sosednjih odsekov odločamo, ali postavimo mejo med segmentoma ali ne. Najbolj uspešne mere podobnosti, ki se uporabljajo pri takšni segmentaciji, so simetrična Kullback–Leiblerjeva (divergenčna, KL2) mera verjetnostne podobnosti [Siegler-97], mera LLR [Bonastre-00, Delacourt-01, Mori-01, Ajmera-03] in informacijski kriteriji, med njimi najbolj učinkovit Bayesov informacijski kriterij, BIC, [Chen-98, Delacourt-01, Mori-01,

Tritschler-99, Lopez-00, Zhou-00, Vandecatseye-03, Cettolo-05]. V postopkih segmentacije na različne načine določamo meje med segmenti v tistih točkah, kjer s posameznimi merami dosežemo lokalne maksimume ali minimume.

Podobne metode segmentacije se uporabljajo tudi na drugih raziskovalnih področjih, kjer je potrebno daljša zaporedja podatkov ali osnovnih enot razdeliti na manjše odseke. Tako se npr. pri analizi DNA vijačnic uporablja segmentacija s sprotnim odločanjem [Li-02], kjer mere segmentacije v glavnem temeljijo na meri LLR oziroma na informacijskih kriterijih.

Vsaka izmed teh metod ima svoje prednosti in pomanjkljivosti. Pri metodah segmentacije s predhodnim učenjem izvajamo segmentacijo z modeli, ki jih predhodno določimo na podlagi učnega materiala. Tako se lahko zgodi, da z njimi ne moremo dovolj dobro opisati spremenjenih akustičnih razmer in situacij, ki jih z učnimi podatki nismo uspeli zajeti. To pa posledično pomeni slabšo segmentacijo posnetkov v takšnih primerih. Podobno je tudi v primeru postopkov segmentacije na podlagi mer podobnosti ali različnosti. Tu določamo meje med segmenti na podlagi pragov odločitve, ki jih moramo predhodno določiti. Za to običajno potrebujemo dodatne razvojne zbirke podatkov, kjer na podlagi optimalnih rezultatov segmentacije nastavljamo pragove odločitve. Bistvena prednost takšnih postopkov je ravno v tem, da za določitev pragov potrebujemo manj ustrezno označenih akustičnih podatkov kot pri učenju modelov segmentacije. Zato se te postopke uporablja predvsem v sistemih predhodne obdelave zvočnih posnetkov, kjer je potrebno parametre postopkov prilagajati trenutnim razmeram delovanja.

Izkazalo se je tudi, da s postopki segmentacije s sprotnim odločanjem dobimo boljše rezultate segmentacij po govorcih [Kemp-00]. Zato smo se tudi mi v raziskovalnem delu omejili predvsem na preučevanje te skupine postopkov. V nadaljevanju bomo tako podali pregled dosedanjega dela na tem področju.

Največji napredek pri segmentaciji zvočnih posnetkov sta dosegla Chen in Gopalakrishnan [Chen-98], ki sta formulirala problem segmentacije po govorcih kot problem izbire med modeli. Tako sta vsak segment predstavila z enimi modelom, kjer sta ugotavljala, ali dva posamezna segmenta boljše opišemo z dvema modeloma ali z enim skupnim. Pri tem sta za primerjave med modeli segmentov vpeljala mero BIC. Osnovnemu kriteriju BIC, ki je bil prvič definiran v [Scwartz-76], sta dodala še en utežni faktor (običajno označen z λ), s katerim sta implicitno definirala prag odločitve za meje. Utežni faktor sta določala na podlagi razvojnih zbirk. V številnih eksperimentih [Tritschler-99, Kemp-00, Delacourt-01, Mori-01, Vandecatseye-03, Lopez-00, Cettolo-00, Ajmera-04, Žibert-05] se je izkazalo, da s pravo izbiro faktorja uteži znatno izboljšamo rezultate segmentacije. Osnovni postopek segmentacije s kriterijem BIC sta izboljšala Tritschler in Gopinath [Tritschler-99], ki sta vpeljala številne pohitritve postopka Chena in Gopalakrishnana. Dodatno sta se ukvarjala tudi z detekcijo kratkih segmentov (dolžine manj kot 2 s), ki jih pri segmentaciji s kriterijem BIC težko modeliramo. Podobne izboljšave so bile predlagane tudi v [Cettolo-05], kjer so z različnimi načini ocenjevanja kovariančnih matrik Gaussovih porazdelitev, ki nastopajo v kriteriju BIC, znatno pohitrili postopke in izboljšali rezultate segmentacije. Drugačen pristop pohitritve postopkov določanja mej med segmenti s kriterijem BIC je bil predlagan v

postopku DISTBIC [Delacourt-01]. Osnovna ideja je bila, da bi s hitrim postopkom najprej določili kandidate za meje med segmenti, potem pa bi na podlagi kriterija BIC izbrali prave meje. V tem primeru je segmentacija potekala v dveh fazah, v prvi fazi se je z različnimi merami podobnosti določalo kandidate za meje, v drugi pa s kriterijem BIC izbiralo med njimi. V primeru [Delacourt-01] so v prvi fazi uporabljali mero razmerja logaritma verjetnostne ocene levega in desnega odseka *(ang. log–likelihood ratio, LLR)*, v [Zhou-00] pa so predlagali uporabo T^2 statistike za določanje kandidatov.
Zanimiv eksperiment segmentacije s kriterijem BIC so izvedli tudi v [Vandecatseye-03], kjer so z uporabo normaliziranih vrednosti ocen kriterija BIC dosegli boljše rezultate segmentacije kot v primeru ne–normaliziranih ocen.

Kljub temu, da s kriterijem BIC dosegamo najboljše rezultate segmentacije in se ga zato skoraj izključno uporablja pri segmentaciji zvočnih posnetkov po govorcih in/ali spremembah akustičnega ozadja, je bilo veliko poskusov narejenih tudi z drugimi merami podobnosti ali različnosti. Največkrat se je v teh primerih uporabljala razdalja KL2 [Siegler-97, Žibert-05]. Obstajajo pa tudi druge mere. Tako je Gish s sod. [Gish-91] predlagal svojo mero, ki je temeljila na meri LLR in je bila uporabljena tudi pri segmentaciji v [Kemp-00]. Mori in Nakagawa [Mori-01] sta predlagala za mero kriterij popačenja vektorske kvantizacije *(ang. vector quantisation distortion criterion)* in ga primerjala s kriterijem BIC in mero LLR.
V študiji, ki jo je opravil Kemp s sod. [Kemp-00], so primerjali različne metode segmentacije. Ugotovili so, da z metodami, ki temeljijo na predhodnem učenju modelov, zelo natančno določimo dejanske meje med segmenti, vendar pri tem detektiramo preveč mej. Ravno obratno pa je veljalo za postopke segmentacije s sprotnim odločanjem, kjer zelo dobro ocenimo število dejanskih mej med segmenti, vendar so te nenatančno postavljene, zato so predlagali kombinacijo obeh principov segmentacije. Do podobnih ugotovitev sta prišla tudi Liu in Kubala [Liu-99], ki sta predlagala nov kriterij določanja mej na podlagi samodejno pridobljenih transkripcij razpoznavanja govora.

V vseh omenjenih postopkih in predlaganih kriterijih je potrebno na nek način določati prag odločitve za mejo. To pa predstavlja največji problem postopkov segmentacije. Zato smo se v disertaciji posvetili postopkom, s katerimi bi v največji možni meri zmanjšali vpliv pragov odločitve na rezultate segmentacije v različnih pogojih delovanja.

Omeniti moramo še, da smo se v okviru raziskovalnega dela doktorske disertacije omejili samo na segmentacijo zvočnih posnetkov, vendar se lahko v kombinaciji z video signalom izvaja tudi t.i. bimodalna segmentacija posnetkov. Ta je bila uspešno izvedena v primeru segmentacije po govorcih [Iyengar-00] ter v primeru detekcije in sledenja posameznim vsebinam televizijskih posnetkov informativnih oddaj [Iurgel-01].

1.3.2.3 Razvrščanje segmentov po govorcih s postopki rojenja

Pri razvrščanju segmentov po govorcih gre v bistvu za določanje segmentov, ki pripadajo istim govorcem. Določanje pripadnosti segmentov k govorcem pa izvajamo s postopki rojenja.

Večina najbolj uspešnih sistemov s tega področja temelji na hierarhičnih postopkih

združevanja segmentov v roje. To pomeni, da segmente, ki jih pridobimo pri segmentaciji zvočnih posnetkov po govorcih, združujemo v skupne roje toliko časa, dokler z vsakim rojem ne opišemo natanko enega govorca v zvočnem posnetku. Običajna strategija rojenja temelji na združevanju segmentov, ki so si med seboj blizu glede na mero podobnosti, ki jo uporabljamo. Pri tem moramo določiti še, kdaj se z združevanjem ustavimo. Tako se pri razvrščanju segmentov po govorcih ukvarjamo predvsem s predstavitvami segmentov, ki bi bile primerne za združevanje po govorcih, z merami podobnosti za združevanje in s kriteriji zaustavitve rojenja. V nadaljevanju bomo podali nekaj temeljnih del s tega področja.

Eden izmed prvih poskusov združevanja segmentov po govorcih, ki ga je predlagal Jin s sod. [Jin-97], je bil namenjen prilagajanju govornih modelov na posameznega govorca *(ang. speaker adaptation)* v sistemu za razpoznavanje govora. Tu je bil vsak segment predstavljen z modelom ene Gaussove porazdelitve, združevanje segmentov pa je potekalo z mero podobnosti, ki je bila predlagana v [Gish-91]. Kriterij zaustavitve je bil izpeljan kot kombinacija povprečne razdalje med pari segmentov znotraj posameznega roja *(ang. within–cluster dispersion)* na eni strani in faktorja kaznovanja, ki je bil odvisen od trenutnega števila rojev, na drugi strani. Pogoj za zaustavitev rojenja je bil dosežen minimum kriterija zaustavitve. Faktor kaznovanja je bil nujen, saj bi sicer minimum kriterija zaustavitve dosegli v primeru, kjer bi vsak roj vključeval samo en segment. Zato so izvajali številne preizkuse z različnimi izbirami faktorjev kaznovanja, vendar sistematičnih rešitev za določitev kriterija zaustavitve niso podali. Pokazali pa so, da z združevanjem segmentov v roje glede na govorce in s prilagajanjem modelov govora na ta način znatno izboljšamo rezultate razpoznavanja govora.

Drugi način rojenja segmentov je bil predstavljen v [Siegler-97], kjer je združevanje segmentov potekalo glede na Kullback–Leiblerjevo (KL) razdaljo med segmenti. V postopku rojenja je združevanje segmentov v roje potekalo tako, da sta bila dva segmenta ali roja (odvisno od koraka rojenja) združena, če je bila njuna medsebojna KL razdalja pod določenim pragom združevanja, ki so ga podali vnaprej. In ravno ta prag združevanja predstavlja največji problem te metode. Zato so bile preizkušane številne tehnike določanja praga, vendar sistematičnih rešitev ni bilo predlaganih.

Podobno mero podobnosti za združevanje so predlagali tudi v [Solomonoff-98], kjer je bil za razvrščanje segmentov po govorcih uporabljen hierarhičen postopek rojenja s pomočjo dendrogramov. Dendrogrami združevanja so bili zgrajeni na podlagi dveh mer, mere LLR in razdalje KL. Iskanje optimalnega števila rojev pa je bilo izvedeno s pomočjo kriterija največje čistosti rojev *(ang. cluster purity)*, ki so ga prav tako predlagali v tem delu.

Drugačen način razvrščanja segmentov po govorcih je bil predlagan v [Johnson-99]. Postopek rojenja, ki so ga preizkušali, je prav tako temeljil na postopkih hierarhičnega rojenja, vendar so tu uporabljali kombinacijo rojenja od zgoraj–navzdol s postopki združevanja. Ideja postopka je bila predvsem vezana na metodo prilagajanja modelov po govorcih v sistemih za razpoznavanje govora z MLLR[6] adaptacijo [Johnson-98]. V tem primeru so za deljenje in za združevanje rojev uporabljali dve meri, mero AHS[7]

[6] MLLR je kratica za postopek Maximum Likelihood Linear Regression, [Gales-96].
[7] AHS je kratica za mero Arithmetic Harmonic Sphericity.

[Bimbot-93] in mero Gaussove divergence. V prvi fazi se je rojenje izvajalo od zgoraj–navzdol, v drugi pa je potekalo združevanje tistih rojev, ki so se preveč delili. Kriteriji za zaustavitve postopkov deljenja in združevanja so bili določeni tako, da so z njimi maksimizirali ocene Gaussovih verjetnostnih porazdelitev na danih podatkih, kar je bilo primerno za MLLR adaptacijo. Ta postopek razvrščanja segmentov po govorcih je bil vključen v sistem za samodejno pridobivanje transkripcij informativnih oddaj [Hain-98, Woodland-02].

Ker sta si problema segmentacije po govorcih in združevanja segmentov po govorcih zelo podobna, saj se v prvem primeru odločamo, ali bomo združili dva segmenta (ne bomo postavili meje) ali jih pustili razdružena (bomo postavili mejo), v drugem pa poteka združevanje več segmentov skupaj, se je tudi tu za najbolj učinkovito mero združevanja izkazal kriterij BIC. Ravno tako kot pri segmentaciji, sta ga tudi pri razvrščanju segmentov prva uporabila Chen in Gopalakrishnan [Chen-98]. Na ta način sta vsak segment predstavila z enim modelom, s postopkom rojenja pa sta na vsakem koraku združevala tiste segmente ali roje, kjer so bile s kriterijem BIC dosežene maksimalne vrednosti. Postopek združevanja je bil končan, ko se je s kriterijem BIC presegel določen prag združevanja, ali ko s kriterijem ni bilo več mogoče povečati BIC vrednosti. Tudi tu, se je podobno kot v primeru segmentacije, vpeljal dodaten utežni faktor λ, s katerim je bilo mogoče nadzirati potek združevanja. V številnih eksperimentih [Tritschler-99, Lapidot-03, Vandecatseye-03, Žibert-05] se je izkazalo, da je potrebno faktor λ dodatno nastavljati in popravljati v primeru različnih akustičnih pogojev v zvočnih posnetkih.

Medtem ko je bilo v večini omenjenih primerov razvrščanje segmentov po govorcih uporabljeno za namene prileganja govornih modelov v sistemih za razpoznavanje govora, smo se v disertaciji osredotočili k rojenju segmentov zvočnih posnetkov za namene indeksacije. V tem primeru smo postopke rojenja in segmentacije izvajali z namenom strukturiranja informativnih oddaj glede na prisotnost govorcev v zvočnih posnetkih. Tako označevanje zvočnih posnetkov je bilo prvič predstavljeno v okviru projekta *Rich Transcription* [Fiscus-05]. Tu so v evaluacijah *"Who spoke when"* preizkušali številne postopke segmentacije in združevanja segmentov po govorcih. Pri večini postopkov so za glavno mero podobnosti med segmenti uporabljali kriterij BIC [Nguyen-03, Moraru-03b, Reynolds-05]. V zadnjih evaluacijah pa so se pojavili že prvi sistemi, kjer je združevanje segmentov potekalo z metodami, ki se uspešno uporabljajo pri razpoznavanju govorcev. Prvi tak sistem so predlagali v [Barras-04], kjer so segmente predstavili z GMM modeli, ki so jih pridobili z MAP adaptacijo splošnih modelov govora *(ang. universal background models, UBM)* [Reynolds-95]. Za mero združevanja so uporabili podoben kriterij, kot se uporablja v sistemih za verifikacijo govorcev. V nadaljnjih raziskavah [Zhu-05, Sinha-05] so preizkušali še s številnimi tehnikami normalizacije akustičnih predstavitev in z izboljšavami postopkov pridobivanja GMM modelov.

V okviru disertacije smo razvili in preizkušali referenčni postopek, predstavljen v [Chen-98], postopek, ki je temeljil na uporabi metod iz razpoznavanja govora [Barras-04, Zhu-05, Sinha-05], kjer smo preizkusili različne kriterije združevanja, razvili pa smo tudi nov postopek, kjer smo segmente združevali na podlagi akustične in prozodične informacije.

1.4 Pregled vsebine disertacije

Disertacija obsega šest poglavij in en dodatek. V uvodnem poglavju smo natančnejše opredelili raziskovalno področje in naloge, s katerimi smo se ukvarjali v okviru doktorskega dela. Ker raziskovalno področje teme disertacije vključuje številna področja uporabe, so tudi naloge, ki jih rešujemo, različne. Zato smo najprej podali pregled širšega področja obdelave zvočnih posnetkov informativnih oddaj z uporabo govornih tehnologij ter vanje ustrezno umestili naloge, s katerimi smo se ukvarjali v našem raziskovalnem delu. V pregledu ožjega področja teme disertacije pa smo podali pregled temeljnih del, s katerimi se bomo ukvarjali v naslednjih poglavjih.

V drugem poglavju bomo tako najprej predstavili dve podatkovni zbirki posnetkov informativnih oddaj, ki smo ju uporabljali pri razvoju in vrednotenju postopkov segmentacije in razvrščanja segmentov zvočnih posnetkov. Prva zbirka predstavlja zbirko informativnih oddaj v slovenskem jeziku in je bila pridobljena v okviru raziskovalnega dela doktorske disertacije. Namenjena je predvsem razvoju splošnega sistema za razpoznavanje govora v slovenskem jeziku. Drugo zbirko smo pridobili v okviru sodelovanja v mednarodnem projektu COST278 in predstavlja prvo večjezično zbirko posnetkov informativnih oddaj, namenjeno razvoju postopkov obdelave informativnih oddaj, ki so neodvisni od jezika. V tem poglavju je opisan postopek pridobivanja posnetkov informativnih oddaj, proces označevanja zvočnih posnetkov, podana pa je tudi analiza in primerjava obeh zbirk.

Glavne naloge, ki smo jih reševali v disertaciji, so opisane v poglavjih 3, 4 in 5. V tretjem poglavju se ukvarjamo s segmentacijo zvočnih posnetkov na govorne in ne–govorne dele, v četrtem s segmentacijo glede na zamenjave govorcev in spremembe v akustičnem ozadju, peto poglavje pa je namenjeno razvrščanju segmentov po govorcih s postopki rojenja.

Problem detekcije govora v zvočnih posnetkih v tretjem poglavju predstavimo kot problem segmentacije zvočnih posnetkov na govorne in ne–govorne odseke. Zato se v tem poglavju ukvarjamo v glavnem s primernimi predstavitvami zvočnih signalov za detekcijo govora in s postopki segmentacije na podlagi teh predstavitev. Predlagane so štiri nove značilke, ki jih pridobivamo neposredno iz transkripcij samodejnega razpoznavanja glasov, in nov postopek segmentacije na podlagi teh značilk. Primerjava standardnih - akustičnih in predlaganih - fonetičnih predstavitev zvočnih posnetkov ter različnih postopkov segmentacije je bila izvedena na podatkovnih zbirkah iz drugega poglavja in je prav tako predstavljena v tem poglavju.

V četrtem poglavju se ukvarjamo predvsem s segmentacijo zvočnih posnetkov glede na zamenjave govorcev. Tu smo izboljšali standardni postopek segmentacije s kriterijem BIC tako, da ni več potrebno določati praga odločitve za meje med segmenti vnaprej, ampak se prag odločitve določa relativno glede na akustične razmere v obdelovanih posnetkih. Tako sta v četrtem poglavju predlagani dve metodi segmentacije: postopek segmentacije z relativno določenim pragom in postopek segmentacije z združevanjem različnih predstavitev zvočnih posnetkov.

Peto poglavje je namenjeno razvrščanju govornih segmentov k istim govorcem. To do-

sežemo s postopki rojenja z združevanjem, kjer se ukvarjamo s predstavitvami govornih segmentov, primernimi za združevanje po govorcih, z merami združevanja in kriteriji zaustavitve rojenja. Primerjamo tri postopke rojenja: osnovni postopek s kriterijem BIC in akustičnimi predstavitvami segmentov, postopek, ki temelji na uporabi metod iz razpoznavanja govora, in nov postopek, ki smo ga razvili iz osnovnega postopka z dodajanjem prozodične informacije. Vrednotenje postopkov je izvedeno v primeru idealne segmentacije, kjer so bili segmenti označeni ročno in v primeru samodejne segmentacije, kjer je bila segmentacija izvedena s postopki iz prejšnjih poglavij. V zadnjem delu poglavja se posvečamo kriterijem zaustavitve rojenja. Predlagamo dva nova kriterija in primerjamo rezultate razvrščanja na podlagi teh kriterijev. Ti rezultati predstavljajo hkrati tudi končne rezultate samodejnega označevanja informativnih oddaj glede na prisotnost govorcev v zvočnih posnetkih.

V zaključnem poglavju najprej povzamemo in poudarimo bistvene prispevke, ki smo jih predlagali za izboljšanje postopkov segmentacije in razvrščanja segmentov. V nadaljevanju pa obravnavamo možne izvedbe predlaganih postopkov in vključevanje v različne sisteme govornih tehnologij, kjer se osredotočimo na izvedbo sistema za samodejno indeksacijo zvočnih posnetkov informativnih oddaj.

V dodatku A podrobneje opišemo razdelitev zvočnih posnetkov informativnih oddaj iz zbirk SiBN in COST278 na učne, razvojne in testne množice, ki smo jih uporabljali v posameznih preizkusih postopkov iz disertacije.

2 Podatkovne zbirke zvočnih posnetkov informativnih oddaj

V tem poglavju bomo predstavili in primerjali dve podatkovni zbirki posnetkov informativnih oddaj: slovensko zbirko SiBN in večjezično zbirko informativnih oddaj COST278. Zbirka SiBN je bila namensko pridobljena za potrebe doktorske disertacije in je označena tako, da jo bomo lahko uporabili tudi za izgradnjo različnih sistemov za samodejno razpoznavanje govora. Poleg posnetkov informativnih oddaj smo pridobili in ustrezno pripravili tudi jezikovni korpus informativnih oddaj, ki je pridružen osnovni zbirki SiBN. Večjezična jezikovna zbirka informativnih oddaj COST278 pa je bila pridobljena v okviru mednarodnega sodelovanja v projektu COST278. Sestavljena je iz usklajeno označenih zvočnih posnetkov informativnih oddaj v devetih evropskih jezikih in je primerna za razvoj in vrednotenje postopkov govornih aplikacij, ki so neodvisne od jezika. Zaradi raznolikosti zbranih posnetkov jo lahko uporabljamo tudi za vrednotenje neobčutljivosti postopkov v različnih pogojih delovanja. V okviru doktorske disertacije smo ravno zaradi specifičnih lastnosti obeh podatkovnih zbirk izvajali eksperimente na obeh zbirkah. S tem smo lahko bolj objektivno ocenjevali in primerjali učinkovitost posameznih postopkov v različnih pogojih delovanja.

2.1 Podatkovne zbirke informativnih oddaj

Ker so informativne oddaje v večini primerov sestavljene iz različnih multimedijskih vsebin, so tudi podatkovne zbirke informativnih oddaj sestavljene iz različnih tipov podatkov, ki jih je potrebno označiti in uskladiti med seboj. Kakšne podatke in koliko jih potrebujemo, pa je odvisno od sistemov, ki jih načrtujemo, in nalog, ki jih rešujemo.

Tako se v primeru podatkovnih zbirk informativnih oddaj, ki jih uporabljamo za razvoj sistemov govornih tehnologij oziroma širše jezikovnih tehnologij, omejimo na dva osnovna tipa podatkov: govorne in tekstovne podatke. Govorne podatke pridobivamo predvsem na podlagi zvočnih posnetkov, ki jim po potrebi (če je mogoče) pridružimo še video posnetke. Zbrane posnetke ustrezno označujemo in usklajujemo med seboj. Stopnja označevanja posnetkov je odvisna od namena in uporabe zbirke. Običajno se v takšnih zbirkah poleg osnovnih tekstovnih prepisov govora označuje še značilnosti govorca, tip govora, akustična ozadja, vsebino govora ipd. Zato v tem primeru govorimo, da smo osnovno informacijo tekstovnega prepisa govora obogatili z dodatno govorno informacijo *(ang. rich transcriptions)*. Skupaj z govorno informacijo običajno takšne zbirke vsebujejo še tekstovne podatke, ki so predstavljeni z velikimi korpusi besedil sorodnih vsebin, kot so vsebine, ki so zajete v govornih posnetkih informativnih oddaj. Oboje skupaj lahko štejemo za podatkovno zbirko informativnih oddaj *(ang. broadcast news speech database)*, ki jo uporabljamo kot govorno zbirko za razvoj aplikacij govornih tehnologij.

Posebnost takšnih govornih zbirk je v tem, da niso načrtovane in se močno razlikujejo od namensko pridobljenih govornih zbirk. Osnovne razlike lahko strnemo na naslednja področja:

- *akustične lastnosti posnetkov:* govorni posnetki v namenskih zbirkah so pridobljeni običajno v nadzorovanih akustičnih pogojih, posnetki informativnih oddaj pa zaradi narave pridobivanja novic v različnih akustičnih pogojih;
- *tip govora:* v namenskih zbirkah je običajno zajet en tip govora (bran govor, ločeno izgovorjene besede,...), v informativnih oddajah je prisotno več tipov govora (bran, spontan, čustven, govor, ne–govor, ...);
- *govorci:* v namenskih zbirkah imamo običajno manjše število govorcev, ki so enakomerno zastopani, nimamo tujih govorcev; pri zbirkah informativnih oddaj imamo večje število govorcev, ki so zelo neenakomerno zastopani, prisotna je različna kvaliteta govorne artikulacije, imamo tudi znatno število tujih govorcev ipd.;
- *jezik:* v namenskih zbirkah imamo običajno tematsko omejene vsebine govora, kar posledično pomeni sorazmeroma majhna besedišča in temu primerno dobre jezikovne modele govora; v primeru informativnih oddaj imamo različne vsebine, posledično velika besedišča, tuj govor, različne dialekte, ...

Ravno zaradi kompleksnosti podatkov in zaradi velikih količin različnih informacij, ki so vsebovane v govornih in tekstovnih podatkih, predstavljajo podatkovne zbirke

informativnih oddaj velik izziv za zbiranje, označevanje in usklajevanje različnih tipov podatkov, ki jih lahko pridobivamo iz informativnih oddaj. Po drugi strani ravno zaradi teh lastnosti in možnosti sprotnega razširjanja zbirk z novimi podatki predstavljajo zbirke informativnih oddaj skoraj neomejen vir podatkov, primernih za izgradnjo in razvoj sistemov z različnih področij govornih tehnologij. Tipično govorne podatke za takšne zbirke pridobivamo iz radijskih in televizijskih informativnih oddaj, tekstovni material pa iz dnevnega časopisja oziroma preko informacijskih portalov iz medmrežja.

Pridobivanje in označevanje večjih govornih podatkovnih zbirk informativnih oddaj se je koncem devetdesetih let prejšnjega stoletja začelo z namenom, da bi pridobili dovolj govornih podatkov za izgradnjo večjih sistemov za razpoznavanje govora, ki bi bili primerni za razpoznavanje splošnega govora v različnih akustičnih pogojih. Prva uporaba takšnih sistemov je bila usmerjena k samodejnemu pridobivanju tekstovnih prepisov zvočnih posnetkov informativnih oddaj [Woodland-02, Beyerlein-02, Chen-02, Gauvain-02]. Z razpoznavanjem in vključevanjem dodatne informacije osnovnim (samodejnim) tekstovnim prepisom govora je bil omogočen tudi razvoj novih sistemov za samodejno pridobivanje, sledenje in indeksacijo različnih vsebin informativnih oddaj [Wayne-00], ki smo jih pregledali že v uvodnem poglavju. Osnova za razvoj različnih postopkov obdelave zvočnih posnetkov, ki so vključene v takšne sisteme, so pravilno pridobljene in ustrezno označene podatkovne zbirke. Zato smo tudi mi začeli z razvojem in pridobivanjem takšnih zbirk.

V nadaljevanju bo tako predstavljena podatkovna zbirka informativnih oddaj v slovenskem jeziku – SiBN in večjezična zbirka COST278. V okviru doktorske disertacije smo obe zbirki uporabljali za razvoj in vrednotenje postopkov detekcije govora, segmentacije in razvrščanja segmentov po govorcih za namene indeksacije zvočnih posnetkov. Medtem ko je bila zbirka SiBN zasnovana širše za razvoj LVCSR sistemov, je bila zbirka COST278 namenjena ravno postopkom, s katerimi smo se ukvarjali v okviru našega raziskovalnega dela. Ker sta bili obe zbirki enako zasnovani, bomo tako najprej predstavili, kako smo zajemali in označevali posnetke informativnih oddaj in katera orodja smo pri tem uporabljali, v nadaljevanju pa bomo podali podrobnejšo analizo obeh zbirk.

2.1.1 Pridobivanje podatkovnih zbirk informativnih oddaj

V tem razdelku bomo opisali skupne značilnosti obeh podatkovnih zbirk, opisali potek označevanja zvočnih posnetkov ter predstavili orodja za pretvorbo med različnimi formati transkripcij in za preverjanje transkripcij, ki smo jih razvili vzporedno z označevanjem podatkov.

Kljub temu da sta bili obe zbirki zasnovani za različne namene, je osnovna struktura obeh zbirk enaka. Obe zbirki vsebujeta zvočne in video posnetke samo televizijskih informativnih oddaj iz različnih televizijskih postaj, ki so bile posnete v različnih obdobjih. Zbirka SiBN je bila zasnovana širše in poleg govornega materiala vključuje še tekstovni material, ki smo ga pridobili z zbiranjem besedil za podnaslavljanje informativnih oddaj preko teleteksta.

Ker je razvoj obeh zbirk potekal skoraj sočasno, smo pri obeh zbirkah uporabili enaka pravila za pridobivanje in označevanje zvočnih posnetkov ter za usklajevanje z video posnetki. Glavna razlika je bila v namenu uporabe obeh zbirk. Zbirka COST278 je bila zasnovana tako, da je vsaka partnerska skupina, ki je sodelovala pri izdelavi zbirke, prispevala 3 ure posnetkov informativnih oddaj v svojem jeziku in je bila tako namenjena predvsem razvoju postopkov obdelave informativnih oddaj, ki so neodvisne od jezika. Skupaj smo tako dobili 30 ur posnetkov v devetih jezikih. Zbirka SiBN pa trenutno vsebuje 34 ur posnetkov informativnih oddaj v slovenskem jeziku in je namenjena predvsem izgradnji splošnega sistema za razpoznavanje govora v slovenskem jeziku. Druga pomembna razlika med zbirkama je tudi v raznolikosti posnetkov. V zbirki COST278 so zbrani posnetki informativnih oddaj različnih TV postaj, kar posledično pomeni večjo vsebinsko in strukturno pestrost zbranih posnetkov. Posnetki zbirke SiBN pa vsebujejo informativne oddaje samo ene TV postaje, kar pomeni, da so podatki bolj homogeni in posnetki podobno strukturirani. Tako lahko zaključimo, da so v zbirki SiBN zbrani podatki jezikovno, vsebinsko in strukturno bolj homogeni, namen zbirke COST278 pa je bil ravno nasproten, in sicer uskladiti in poenotiti transkripcije informativnih oddaj iz različnih zbirk, ki bi bile jezikovno, vsebinsko in strukturno čim bolj raznolike.

Ravno zato smo v vseh naših eksperimentih uporabljali obe zbirki: zbirko SiBN za primerjanje postopkov v primeru bolj homogenih posnetkov, zbirko COST278 pa za testiranje neobčutljivosti postopkov v različnih pogojih delovanja.

2.1.2 Označevanje zvočnih posnetkov informativnih oddaj

Bistveni element vsake take podatkovne zbirke so ustrezne transkripcije[1] zvočnih posnetkov informativnih oddaj. Proces pridobivanja transkripcij zvočnih posnetkov imenujemo označevanje zvočnih posnetkov. Za razliko od namenskih govornih zbirk je v primeru informativnih oddaj označevanje posnetkov najbolj zahtevno opravilo, saj je potrebno posnetke označiti in dokumentirati na različnih nivojih, uskladiti oznake za različne (ne)govorne in (ne)jezikovne pojave in zaradi velike količine podatkov organizirati označevanje in preverjanje transkripcij v več fazah.

Običajno transkripcije zvočnih posnetkov informativnih oddaj, ki jih zbiramo v podatkovnih zbirkah, vključujejo naslednje elemente označevanja:

struktura oddaje: Osrednje informativne oddaje so običajno sestavljene iz več ločenih sklopov novic, ki se nanašajo na določena področja informacij: politične novice (mednarodne, notranje–politične) lokalne novice (regijske informacije, črna kronika ipd.), finančne, kulturne, športne novice, vremenske napovedi ipd. Vsaka informativna oddaja je tako lahko sestavljena iz več ločenih pododdaj oziroma iz ene oddaje, ki je osredotočena samo na določeno informativno področje. Prehodi iz enega sklopa novic k drugemu so običajno povezani z značilnimi avdio-

[1] Transkripcije zvočnih posnetkov informativnih oddaj ne vključujejo samo tekstovnega prepisa govora, ampak še druge oznake, ki jih je potrebno pridobivati v procesu označevanja in bodo predstavljeni v nadaljevanju. Transkripcije zvočnih posnetkov informativnih oddaj v tem poglavju tako pomenijo v celoti označene posnetke informativnih oddaj in ne zgolj tekstovnih prepisov govora.

vizualnimi efekti, ki najavljajo naslednjo oddajo oziroma zaključujejo prvo. Pravimo jim televizijske špice. Znotraj posameznih sklopov novic so zaporedoma nanizana poročila o novicah, ki so lahko dodatno opremljena s komentarji, intervjuji, anketami, lahko pa gre tudi za soočenja več govorcev o določeni temi. Vsako novico tako lahko opremimo z informacijo o tipu novice (poročilo, komentar, intervju, anketa, soočenje), o času in kraju dogajanja ter o akterjih in o dogajanju novice. Dodatno lahko novice razvrstimo še, h kateremu sklopu novic pripadajo. Označevanje novic mora biti usklajeno, pri tem pa običajno sledimo vnaprej pripravljenim ali pridobljenim strukturam novic, ki so značilne za določeno informativno oddajo ali TV postajo.

Tako označene novice informativnih oddaj lahko uporabimo za razvoj sistemov za detekcijo in sledenje vsebinam informativnih oddaj *(ang. topic detection and tracking)*.

govorci: Običajno je pri informativnih oddajah struktura govorcev zelo pestra. Informativne oddaje vsebujejo veliko število govorcev, ki prispevajo zelo malo govora, kar je posledica številnih novic in različnih novinarjev, ki poročajo o novicah, ter akterjev novic. Po drugi strani pa imamo nekaj govorcev z izredno veliko količino govora. To so voditelji informativnih oddaj, ki povezujejo, napovedujejo oziroma prebirajo poročila o novicah. Običajno tudi vodijo soočenja in intervjuje v TV studijih. Ker so to običajno šolani govorci in prebirajo že vnaprej pripravljeno besedilo, je njihov govor zelo kvaliteten in jezik izbran. V drugo skupino sodijo novinarji, ki poročajo ali komentirajo posamezne novice. Ti prav tako prispevajo enako kvaliteten govor in jezik, vendar v manjši količini, govorcev pa je več kot je voditeljev. V zadnjo skupino sodijo predvsem govorci, ki niso novinarji, ampak akterji novic. Vsak tak govorec prispeva izredno malo govornih podatkov, tako njihov govor kot jezik pa sta zelo pestra.

Pri označevanju govorcev običajno poleg osnovne identifikacije o govorcih označujemo še spol govorca, kvaliteto govora in jezika posameznega govorca ter kvaliteto posnetkov zvočnih podatkov.

govor in jezik: Govor in jezik sta v informativnih oddajah izredno pestra. To je predvsem posledica strukture govorcev, ki so vključeni v informativne oddaje. Prvo skupino govora prispevajo govorci, ki poročajo in komentirajo novice, drugo pa sestavlja govor akterjev novic. Kot smo že povedali v prejšnjem primeru, novinarji in voditelji prispevajo zelo kvaliteten, ostali govorci pa manj kvaliteten govor in jezik. V prvi skupini imamo tako pretežno bran govor s tekočo izgovorjavo, skoraj brez napak in dobro artikuliran; jezik je izbran brez pogovornih in narečnih besed, stavki so dobro strukturirani. V drugi skupini pa je zajet govor iz realnega sveta, kar pomeni, da imamo tu od branega do povsem spontanega govora, jezik je v glavnem pogovorni, v primeru lokalnih novic je veliko narečnega govora, veliko je negovornih in nejezikovnih elementov, ki so posledica neizkušenosti in nepripravljenosti nastopanja govorcev pred mikrofonom. Tudi struktura jezika je precej prosta, več je odsekanih in novih sestavljenih besed, nedokončanih stavkov ipd. Poleg tega imamo v informativnih oddajah tudi veliko domačih in tujih lastnih imen, novih imen za nove dogodke in stvari, kratic, raznih rezultatov ipd., kar predstavlja veliko težavo za označevanje in jezikovno modeliranje

informativnih oddaj. Dodatna posebnost informativnih oddaj je tudi v tem, da vsebujejo tudi govor v tujih jezikih, ki ga običajno ne označujemo.

Prav zaradi številnih govornih in jezikovnih lastnosti govora informativnih oddaj predstavlja označevanje govora najtežji del v procesu označevanja. Tu poleg osnovnih tekstovnih prepisov govora označujemo še številne negovorne in nejezikovne elemente v govoru, kot so razni medmeti, nedokončane in nove besede, napačne izgovorjave, govorne napake ipd. Da bi zagotovili usklajenost označevanja, je potrebno ravno zaradi teh elementov dodatno preverjati transkripcije in usklajevati vse oznake. Pri transkripcijah govora dodatno označujemo tudi tip govora in kvaliteto jezika. Govor v tujih jezikih pa običajno ne označujemo, ampak ga samo identificiramo.

ostali elementi: Sem spada označevanje vseh ostalih elementov, ki ne sodijo v nobeno izmed zgornjih skupin. Tu predvsem mislimo na označevanje ne–govornih pojavov v zvočnih posnetkih. Sem spada označevanje kvalitete akustičnega ozadja oziroma označevanje akustičnih dogodkov, ki so slišni v ozadju govornih posnetkov, kot so razni šumi ali glasba v ozadju, označevanje prekrivajočega govora ipd.
Poleg tega označujemo tudi odseke televizijskih špic in identificiramo dele informativnih oddaj, ki jih običajno ne označujemo, kot so reklamni bloki, daljši premori, razne prekinitve ipd. Tudi v tem primeru je potrebno zagotoviti usklajeno označevanje vseh teh elementov.

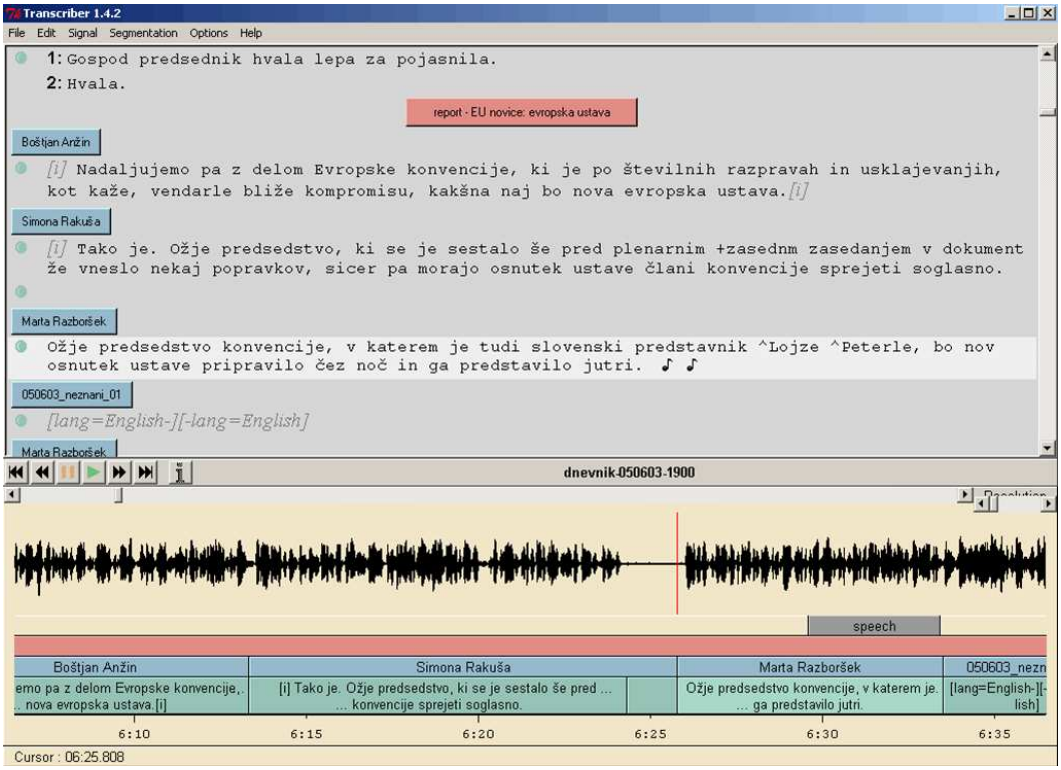
V primeru zbirk SiBN in COST278 smo označevali posnetke informativnih oddaj v skladu s pravili, ki jih je predpisalo združenje LDC[2] [LDC-00] in so jih uporabljali tudi za označevanje posnetkov zbirk v projektu Hub-4 [Graff-02]. V obeh primerih smo označevali vse prej naštete elemente TV informativnih oddaj. Za označevanje smo uporabljali orodje *Transcriber* [Barras-01], ki je prikazano na sliki 2.1.

S tem orodjem smo pridobivali transkripcije zvočnih posnetkov informativnih oddaj, ki so bile organizirane hierarhično glede na naslednje osnovne elemente označevanja:

oddaja *(ang. episode)* : Tu smo označevali osnovne lastnosti informativne oddaje: datum posnetka, ime oddaje, TV postajo in identifikacijo informativne oddaje.

sekcija *(ang. section)* : Sekcija predstavlja vsebinsko enovit del informativne oddaje. Informativna oddaja je običajno razdeljena na več sekcij, ki vključujejo tri skupine vsebin. V prvo skupino spadajo vsi *neoznačeni deli* informativnih oddaj *(ang. notrans)*, kot so reklamna sporočila, televizijske špice *(ang. jingles)* in daljši ne–govorni deli. Drugo skupino predstavljajo vse *najave* v oddaji*(ang. filler)*, ki vključujejo različne preglede vsebin informativnih oddaj, napovedi novic, uvodne in zaključne dele oddaj ipd. Zadnjo skupino pa tvorijo posamezne novice informativnih oddaj. Tu smo označevali osnovni tip novice (poročilo, komentar, intervju, soočenje), področje vsebine (mednarodne, domače, lokalne novice, finance, kultura, šport, ...) ter osnovne podatke o novici (čas in kraj dogajanja,

[2]Lingustic Data Consortium, http://www.ldc.upenn.edu/



Slika 2.1: *Transcriber:* orodje za označevanje zvočnih posnetkov informativnih oddaj, ki smo ga uporabljali pri označevanju posnetkov zbirk SiBN in COST278.

akterje novice, opis dogodka). Pri označevanju novic smo upoštevali hierarhično lestvico razvrstitve področij vsebin, ki smo jo vnaprej določili.

odsek govorca ***(ang. speaker turn)*** : Tu je zajet govor enega samega govorca. Označbe govorca in kvalitete govora bomo opisali v nadaljevanju.

segment ***(ang. segment)*** : Predstavlja osnovni odsek transkripcije. Običajno vključuje odsek govora enega govorca, ki je ločen z daljšim premorom ali vzdihom (izdihom) govorca, ko tvori nov stavek ali del povedi. V redkih primerih pa je v segmentih zajet tudi prekrivajoč govor dveh ali več govorcev. Po kakšnih pravilih smo določali segmente, bo opisano v nadaljevanju.

Osnovni elementi označevanja *odsekov govorcev* so zbrani v tabeli 2.1. Označevanje odsekov govorcev je vključevalo tri glavne skupine označb: elemente govorca, jezik govorca in kvaliteto govora. Dodatno smo označevali tudi akustično ozadje govora pri posameznem govorcu, ki pa se je lahko spreminjalo znotraj posameznega odseka. Običajno je časovna razporeditev in trajanje akustičnih ozadij neodvisno od govora v ospredju, zato je potrebno voditi dva časovna toka označitev: v prvem označujemo govor, v drugem pa spremembe v akustičnem ozadju. Usklajevanje in združevanje obeh tokov označitev predstavlja dodaten problem pri nadaljnji obdelavi transkripcij in ga bomo predstavili v nadaljevanju. Osnovno pravilo pri združevanju je, da se akustična ozadja lahko spreminjajo znotraj enega odseka govorcev, ne morejo pa se raztezati čez enega ali več odsekov. Kot je razvidno iz tabele 2.1, smo označevali 9 osnovnih

Tabela 2.1: Osnovni elementi označevanja odsekov govorcev.

element	oznake
govorec: oznaka govorca	*ime_priimek, neznani_id_nn*
govorec: spol	*moški/ženski/neznani*
jezik: naglas govorca	*<opis>*
jezik: dialekt	*materni/nematerni jezik*
govor: tip govora	*bran/spontan govor*
govor: kvaliteta posnetka	*visoka/srednja/nizka*
govor: kanal posnetka	*studio/telefon*
ozadje: akustično ozadje posnetka	*čisto/glasba/govor/ostalo*

elementov odsekov govorcev. Pri oznakah govorcev smo morali poenotiti identifikacije govorcev v celotni zbirki posnetkov informativnih oddaj. Pri tem smo morali biti pozorni na oznake neznanih govorcev, ki smo jih označevali z *neznani_id_nn*, pri čemer smo morali uskladiti zaporedne številke *nn* govorcev po celi zbirki. Pri spolu govorca smo dodali tudi kategorijo *neznani*, v katero smo vključili govor otrok in starejših ljudi, kar je običajna praksa tudi v ostalih zbirkah [Federico-00, Meinedo-03a]. Pri jeziku govorca smo beležili posebnosti posameznih govorcev. Posebna kategorija je bila namenjena govorcem, ki so govorili v jeziku, ki ni njihov materni jezik[3]. Pri govornih elementih govorca smo ocenjevali tip govora: brani ali spontani govor. Kot bran govor smo označevali govor voditeljev informativnih oddaj in novinarjev, ki so brali poročila oziroma komentarje o novicah. Kot spontan govor pa smo označevali predvsem govor akterjev novic, govor pri soočenjih v informativnih oddajah, intervjuje, govor pri anketah ipd. Kvaliteto posnetkov smo določali glede na pogoje prikazane v tabeli 2.2. Nizka kvaliteta posnetkov se v obeh primerih iz tabele 2.2 nanaša na

Tabela 2.2: Označevanje kvalitete in kanala posnetka v odsekih govorcev.

		kanal posnetka	
		studio frekv. širina > 4kHz	*telefon* frekv. širina < 4kHz
kvaliteta posnetka	*nizka*	velik šum kanala	nerazumljiv govor
	srednja	govor v zunanjem okolju	razumljiv govor
	visoka	studijska kvaliteta govora	čist govor

posnetke, ki so bili pridobljeni v težavnih akustičnih pogojih in je bil zato govor težko razumljiv. Srednja kvaliteta posnetkov je pomenila še vedno slabe akustične pogoje zajema posnetkov, vendar je bil govor na posnetkih dobro razumljiv. To je v primeru studijske kvalitete kanala pomenilo zajemanje posnetkov v ne–studijskih razmerah, v primeru telefonskih posnetkov pa smo s tem opisovali razumljiv govor s prisotnostjo izrazitega šuma kanala. Najvišja kvaliteta oziroma visoka kvaliteta posnetkov se je v obeh kanalih nanašala na posnetke, kjer je bil zajet zelo kvaliteten, čist govor brez

[3]To je bil jezik transkripcije, ki smo ga označevali, ampak govorec ni izhajal iz govornega področja tega jezika.

zaznavnih šumov, ki bi bili posledica delovanja kanala.

Osnovni elementi označevanja so bili *segmenti*. Vsak *odsek govorca* je bil tako sestavljen iz enega ali več segmentov govora govorca iz enega odseka. Osnova za določevanje mej med segmenti so bili premori med govorom, ki nastanejo v glavnem zaradi dihanja govorca med govorjenjem. Pravilo za določitev segmentov v obeh zbirkah je bilo naslednje. Če je premor med govorom trajal manj kot 0.5 s, meje med segmentoma nismo označevali, če je bil premor dolg med 0.5 in 1.5 s, smo označili eno mejo med segmentoma, ki je bila postavljena na sredino področja premora oziroma na začetek ali konec vdiha ali izdiha govorca *(ang. inspiration, expiration event)*. Če je bilo področje premora daljše od 1.5 s, smo označili dve meji med segmentoma (na vsakem koncu področja), vmesno področje pa smo označili kot premor. Segmenti so v glavnem vsebovali stavke govora posameznega govorca in običajno niso bili daljši od desetih sekund. Osnova pri označevanju segmentov so bili tekstovni prepisi govora. Pri tem smo se držali priporočil združenja LDC, ki smo jim dodali še svoja pravila označevanja. Tako smo dodatno označevali še:

- lastna imena z znakom `^` pred imenom,
- kratice z znakom `@` ali z `~`, odvisno od načina izgovorjave; z znakom `@` smo označevali kratice, ki so bile izgovorjene kot ena beseda, z znakom `~` pa kratice, ki so bile črkovane,
- tuje besede ali besedne zveze, ki smo jim dodali oznake o jeziku in po potrebi označili še izgovorjavo,
- števila in številske oznake: vsa števila smo pretvorili v besedne oblike, številske oznake, kot so športni rezultati, fizikalne in matematične količine, finančne vrednosti ipd. smo zapisali, kot so bile izgovorjene,
- ne–besedne elemente: medmete smo poenotili in označevali v oglatih oklepajih [], nepravilno izgovorjene besede smo označili in dodali še izgovorjave, nedokončane ali odrezane besede smo označili z znakoma `+` in `-`, odvisno, ali je bila beseda odrezana na začetku ali koncu,
- ne–govorne elemente: poenotili smo oznake za vse pričakovane ne–govorne pojave (smeh, jok, aplavz, šelestenje papirja, zvoke pri zamenjavi mikrofona in še druge zvoke v ozadju).

Pri označevanju govora v tujem jeziku smo označevali samo elemente govorca in kvalitete govora, nismo pa označevali tekstovnih prepisov govora.

Na sliki 2.1 si tako lahko ogledamo večino elementov označevanja zvočnih posnetkov informativnih oddaj. V orodju *Transciber* je delovno okolje razdeljeno na dva dela: spodnji del je namenjen postavljanju mej med segmenti in odseki govorcev, zgornji del pa je namenjen izdelavi transkripcij. V spodnjem delu lepo vidimo del strukture transkripcije ene informativne oddaje. Prva vrstica pod signalom je namenjena označevanju akustičnega ozadja v posnetkih (v našem primeru imamo večinoma čist posnetek, razen govora v ozadju *(speech)* na enem delu). Druga vrstica označuje sekcijo, ki je v našem

primeru namenjena novici o Evropski uniji. Vrstica pod njo prikazuje odseke govorcev. Tu so prikazane samo identifikacije govorcev, ki so v našem primeru kar imena govorcev. Za vpisovanje vseh ostalih elementov odsekov je potrebno odpreti posebno okno. Zadnja vrstica v spodnjem oknu je namenjena oznakam osnovnih segmentov. Zgornje okno pa je namenjeno predvsem vpisovanju osnovnih elementov transkripcije. Tudi v tem oknu lahko vidimo vse elemente označevanja. Glavni del zavzemajo tekstovni prepisi govora z vsemi dodatnimi elementi, ki smo jih že opisali.

Tu moramo še omeniti, da orodje *Transciber* omogoča izdelavo transkripcij v različnih kodnih formatih. Tako so bile vse govorne transkripcije v zbirkah SiBN in COST278 izdelane v nacionalnih kodnih formatih, kasneje pa so bile pretvorjene v format Unicode (UTF-8).

Proces označevanja posnetkov v obeh zbirkah je potekal v več fazah. Kot smo že omenili, je zbirka COST278 sestavljena iz različnih posnetkov informativnih oddaj izbranih iz osnovnih zbirk posnetkov. Zato je bilo pri zbirki COST278 potrebno samo uskladiti in poenotiti že pripravljene transkripcije. Tako je vsaka skupina, ki je prispevala svoje posnetke, uskladila transkripcije, da so ustrezale predpisom označevanja, ki smo jih že opisali v tem razdelku. Posebna pozornost pa je bila namenjena predvsem izvedbi segmentacije posnetkov in usklajevanju oznak za ne–govorne in ne–besedne elemente.

Pri zbirki SiBN je potekal proces označevanja od začetka. Potrebno je bilo označiti vse potrebne elemente transkripcij. Zato je bilo označevanje posnetkov organizirano v več fazah. V prvi fazi je potekala segmentacija oziroma razdelitev zvočnih posnetkov informativnih oddaj glede na osnovne elemente transkripcije. Tu je bilo potrebno določiti meje med osnovnimi *segmenti*, *odseki govorcev* in *sekcijami*.
V drugi fazi je bilo potrebno izvesti natančne prepise govora z vsemi dodatnimi oznakami, ki smo jih opisali pri označevanju segmentov. To je bil najbolj zahteven del postopka označevanja, saj je bila potrebna velika natančnost označevanja in je bilo potrebno upoštevati veliko pravil za določitev oznak. Običajno je to tudi časovno najbolj potraten del postopka označevanja. Zato smo si v primeru informativnih oddaj zbirke SiBN pomagali z delnimi tekstovnimi prepisi, ki smo jih pridobili preko teleteksta iz besedil namenjenih podnaslavljanju informativnih oddaj. S tem smo znatno pohitrili proces označevanja osnovnih govornih segmentov.
V tretji fazi je bilo potrebno določiti oznake za odseke govorcev in označiti sekcije informativnih oddaj. Identifikacijo govorcev in kategorizacijo vsebin novic je bilo potrebno uskladiti s seznami, ki so bili vnaprej pripravljeni in so se sprotno dopolnjevali. V primeru novih govorcev je bilo potrebno njihovo identiteto ugotoviti iz video posnetkov, sicer je govorec dobil oznako *neznani* in ustrezno identifikacijsko številko. V primeru novih vsebin pa je bilo potrebno določiti oznake novice po pravilih, ki smo jih že opisali. Zadnji fazi sta bili namenjeni preverjanju transkripcij. Četrta faza je bila namenjena preverjanju tekstovnih prepisov in ustreznih oznak vseh elementov označevanja. Peta faza pa je bila namenjena preverjanju usklajenosti oznak znotraj transkripcije posnetka informativne oddaje in usklajevanju oznak med vsemi transkripcijami različnih informativnih oddaj, ki so zajete v zbirki.

Proces označevanja ene ure informativne oddaje zbirke SiBN je tako zahteval približno od 12 do 15 ur dela izkušenih označevalcev, dodatne 2 do 3 ure dela pa je bilo potrebno

vložiti za preverjanje pravilnosti transkripcij.

2.1.3 Orodja za nadaljnjo obdelavo transkripcij zvočnih posnetkov

2.1.3.1 Pretvorba transkripcij v format STM

Transkripcije, ki smo jih pridobili z orodjem *Transcriber*, so zapisane v osnovnem formatu XML. Za nadaljnjo uporabo transkripcij pa je bilo potrebno zapise XML oznak prilagoditi postopkom, ki smo jih razvijali. V okviru projektov *Hub-4* in *Rich Transcriptions* je potekalo vrednotenje postopkov na podlagi formata *STM*, ki ga je predpisala organizacija *NIST*[4] [Graff-02]. Zato smo se tudi mi odločili, da za osnovni razvojni format transkripcij v obeh zbirkah uporabljamo format STM.

Format STM je bolj enostaven od formata XML definiranega z orodjem *Transcriber*. Osnovni element označevanja v formatu STM je govorni odsek posnetka enega govorca, kjer se akustične lastnosti signala ne spreminjajo. S transkripcijami v formatu STM tako razdelimo zvočne posnetke na osnovne segmente, pri katerih označujemo začetek in konec segmenta, identifikacijo in spol govorca, osnovno stanje posnetka in tekstovni prepis govora. Osnovna stanja segmentov opišemo s sedmimi razredi in jim pravimo *F–stanja (ang. focus conditions, F–conditions)*. To so:

F0: V to skupino je vključen čist, bran govor, ki je jezikovno pravilen in posnet v najboljših akustičnih pogojih.

F1: Tu je zbran v glavnem spontan govor prav tako posnet v dobrih akustičnih pogojih.

F2: Sem spadajo telefonski posnetki.

F3: Vključuje govor, kjer je prisotna glasba v ozadju.

F4: Vključuje govor, ki je posnet v slabih akustičnih pogojih ali pa je prisoten izrazit šum v ozadju, ki pa ni glasba.

F5: Vključuje govor govorcev, ki govorijo v osnovnem jeziku transkripcije, vendar to ni njihov materni jezik.

FX: Vključuje preostale tipe govora, ki niso zajeti v prejšnjih skupinah.

Postopek pretvorbe osnovnih oznak *Transcriber* formata v *F–stanja* je podan s shemo 2.2.

Kot lahko vidimo iz postopka s sheme 2.2, gre tu za pretvorbo treh skupin oznak formata XML iz tabele 2.1 (jezikovne in govorne oznake ter oznake akustičnih ozadij) v predpisanih sedem F–stanj. Pretvorba oznak v F–stanja ni enolična, potrebno je obdelati tudi primere, ko imamo takšne oznake, da bi lahko posnetek pripisali dvema ali več F–stanjem. V takih primerih se ponavadi odločimo za stanje **FX**, ni pa nujno.

[4]National Institute of Standards and Technology

```
postavi F-stanje na "F0"
if tip_govora = "spontan"                                    //spontan govor
        F-stanje = "F1"
if tip_kanala = "telefon"                                    //telefonski posnetek
        F-stanje = "F2"
if kvaliteta_posnetka = "nizka"   //govor posnet v slabih akustičnih pogojih
        F-stanje = "F4"
if dialekt = "neizviren"                                     //govor v neizvirnem jeziku
        if F-stanje = "F4"
                F-stanje = "FX"
        else
                F-stanje = "F5"
if prekrivajoč_govor                                         //prekrivajoč govor več govorcev
        F-stanje = "F4"
if tip_ozadja = "glasba"                                     //govor z ozadjem glasbe
        if F-stanje = "F4" or F-stanje = "F5"
                F-stanje = "FX"
        else
                F-stanje = "F3"
else                                                         //pretvorba ostalih tipov ozadja
        if F-stanje = "F5"
                F-stanje = "FX"
        else
                F-stanje = "F4"
```

Shema 2.2: Algoritem pretvorbe osnovnih oznak govora v F–stanja.

V primeru, ko imamo spontan govor, segmente pripišemo stanju **F1**, telefonske posnetke pa označimo z **F2**. Poseben primer predstavlja prekrivajoč govor, ki pripada več govorcem in ga pripišemo k skupini **F4**. V to skupino spadajo tudi vsi posnetki z akustičnimi ozadji, razen posnetkov, ki imajo glasbo v ozadju in jih zato pripišemo stanju **F3**. Če segmenti ne ustrezajo nobeni zgornji predpostavki, ostanejo v stanju **F0**.

Poleg pretvorbe v F–stanja je bistvo postopka predvsem v tem, da združuje oznake dveh časovnih tokov označevanja iz formata XML v eno samo zaporedje oznak. Za to je bilo potrebno definirati dodatna pravila pretvorbe, da bi dobili smiselne transkripcije v formatu STM. Kot smo že omenili, smo imeli v osnovnem formatu dva toka označevanja: osnovne govorne prepise in oznake akustičnega ozadja. Združevanje obeh tokov lahko poteka na več načinov. Osnovno pravilo, ki smo se ga držali v našem postopku, je bilo, da se lahko akustična ozadja spreminjajo samo znotraj enega *odseka govorca* in se ne morejo raztezati čez več govorcev. Na ta način smo definirali ujemanje odsekov v formatu XML in STM na nivoju govorcev. Združevanje obeh tokov oznak znotraj posameznih segmentov je potekalo na dva načina. Pri prvem načinu smo v vsakem odseku govorca postavili toliko mej, kolikor jih je bilo postavljenih v obeh tokovih oznak iz formata XML. Osnovni segmenti formata STM v tem primeru so bili definirani med dvema takima mejama. Če je bilo trajanje segmenta krajše od predpisanega

minimalnega trajanja, smo ga pridružili sosednemu segmentu glede na podobnosti med oznakami segmentov in mu spremenili oznake tako, da so bile enake oznakam časovno daljšega segmenta. Na ta način smo zagotovili enake akustične pogoje posnetka znotraj vsakega osnovnega segmenta formata STM, ki smo mu tako lahko predpisali samo eno F–stanje po algoritmu iz sheme 2.2. V tem primeru predstavljajo problem govorni prepisi. Ker smo osnovne segmente iz formata XML razbili na več manjših odsekov, smo s tem razbili tudi govor, zajet v osnovnih segmentih formata XML. Pravilo, ki smo ga tu uporabljali, je bilo, da smo govorni prepis pridružili največjemu izmed odsekov osnovnega segmenta. Na ta način so postale transkripcije govora neveljavne, saj smo izgubili prvotno informacijo o začetku in koncu trajanja govora. Po drugi strani pa smo na ta način pridobili segmentacijo zvočnih posnetkov glede na F–stanja. Takšen postopek pretvorbe se uporablja tudi v postopku, ki je pridružen orodju *Transcriber*, za pretvorbo v format STM. Takšne transkripcije, kjer so osnovni segmenti definirani z govorci in F–stanji, so primerne za razvoj postopkov segmentacije in razvrščanja segmentov po govorcih, zato smo jih uporabljali tudi pri razvoju postopkov doktorske disertacije.

Drugi način združevanja obeh tokov oznak iz formata XML je primeren predvsem za razvoj sistemov za razpoznavanje govora. Tu prav tako upoštevamo osnovno pravilo združevanja, da se lahko akustična ozadja spreminjajo samo znotraj odsekov govorcev iz formata XML. Združevanje oznak znotraj odsekov pa tu poteka na drugačen način. Tu ohranimo osnovne *segmente* govora iz formata XML in s tem obdržimo celotno informacijo o transkripciji govora. F–stanja pa določimo po postopku iz sheme 2.2 na sledeč način. Če imamo eno F–stanje v celotnem osnovnem segmentu, pustimo stanje nespremenjeno, če je F–stanj več, pa ločimo dva primera: če prevladuje eno F–stanje, označimo segment s tem stanjem; če pa so F–stanja zastopana enakomerno, skupno F–stanje osnovnega segmenta označimo kot **FX**. Na ta način ohranjamo enake segmente v obeh formatih transkripcij, manj natančno pa določimo F–stanja osnovnih segmentov in so zato takšne transkripcije primerne predvsem za razvoj postopkov razpoznavanja govora.

Opisani postopek pretvorbe transkripcij iz formata XML v format STM smo uporabljali za pretvorbo transkripcij posnetkov zbirke SiBN. Uporabljen je bil tudi kot osnovni postopek pretvorbe transkripcij iz zbirke COST278 v skupnem eksperimentu več raziskovalnih skupin, ki so sodelovale pri vrednotenju postopkov segmentacije in razvrščanja segmentov na zbirki COST278 [Vandecatseye-04, Žibert-05].

2.1.3.2 Preverjanje transkripcij z video posnetki

Predstavili bomo še eno orodje, ki smo ga razvili za preverjanje pravilnosti transkripcij. Ker smo v obeh zbirkah poleg zvočnih posnetkov pridobivali tudi video posnetke informativnih oddaj, smo razvili orodje za preverjanje transkripcij na podlagi video posnetkov.

Osnova za razvoj orodja je bila pretvorba transkripcij iz osnovnega formata XML v format SMIL[5]. Ta format je namenjen predvsem združevanju različnih tipov podatkov

[5]The Synchronized Multimedia Integration Language: http://www.w3.org/AudioVideo/

multimedijskih vsebin. Mi smo ga izkoristili za združevanje tekstovne informacije pridobljene iz transkripcij informativnih oddaj z video posnetki teh oddaj. Pri tem smo pretvorili vse elemente transkripcij v različne tekstovne tokove, ki smo jih poravnali z zvočnim signalom video posnetkov. Sinhronizacija transkripcij z video posnetki je bila izvedena na nivoju osnovnih segmentov transkripcij, in sicer tako, da smo izvajali poravnavo detektiranih televizijskih špic iz video posnetkov z označenimi špicami iz transkripcij.

Za pretvorbo transkripcij iz formata XML v format SMIL smo uporabljali enak postopek pridobivanja elementov *Transcriber* formata, ki smo ga razvili za pretvarjanje transkripcij v format STM. Pri tem pa smo lahko za razliko od prejšnjega primera enolično pretvorili vse elemente transkripcije v format SMIL, kar nam je omogočalo tudi enolično pretvorbo v nasprotno smer. Ta format smo zato lahko izkoristili za preverjanje in popravljanje napak v transkripcijah.

Slika 2.3: Preverjanje transkripcij z video posnetki. Prikazano je delovanje video predvajalnika, ki lahko prikazuje multimedijske vsebine v formatu SMIL.

Na sliki 2.3 je prikazana izvedba formata SMIL v primeru preverjanja transkripcij z video posnetki. Osnovno orodje za preverjanje transkripcij je bil v tem primeru video predvajalnik, ki zmore prikazovati multimedijske vsebine v formatu SMIL. V osrednjem oknu poteka predvajanje video posnetka, zgoraj in spodaj glede na osrednje okno pa so razporejena okna za prikazovanje tekstovne informacije, pridobljene iz osnovnih transkripcij. V spodnjem oknu poteka prikazovanje osnovnih govornih segmentov s popolnoma opremljenimi tekstovnimi prepisi govora. Levo zgoraj je okno namenjeno oznakam trenutnih govorčevih lastnosti, ki se spreminjajo glede na oznake iz osnovne transkripcije. Desno zgoraj je okno, ki prikazuje trenutne informacije o akustičnem ozadju zvočnega posnetka. Okno zgoraj na sredini pa je namenjeno prikazovanju informacije o trenutni vsebini novic, ki so bile označene v osnovnih transkripcijah. Po potrebi se lahko poljubno sprehajamo po video posnetkih in s tem tudi po transkripcijah ter tako preverjamo točnost transkripcij z video informacijo.

Preverjanje transkripcij z video posnetki informativnih oddaj se je izkazalo za izjemno

koristno. To pa zato, ker so video posnetki TV informativnih oddaj opremljeni s številnimi dodatnimi informacijami o govorcih, o novicah in drugimi podatki, s katerimi lahko preverjamo in dodatno popravljamo nepravilne oznake transkripcij. Tak način preverjanja se je izkazal za učinkovitega pri preverjanju identitete govorcev, ugotavljanju pravilnosti zapisov tujih lastnih in zemljepisnih imen ter pri preverjanju vzrokov za nastanek različnih akustičnih ozadij.

Običajno je potekalo preverjanje pravilnosti transkripcij z video posnetki v zadnji fazi označevanja in je bilo namenjeno predvsem popravkom in dopolnjevanju oznak, ki jih je bilo samo na podlagi zvočnih posnetkov težko določiti.

2.2 Slovenska zbirka informativnih oddaj SiBN

Zbirka SiBN vključuje informativne oddaje v slovenskem jeziku in je bila zasnovana kot podatkovna zbirka za razvoj sistemov za samodejno podnaslavljanje in samodejno pridobivanje vsebin informativnih oddaj. Tako vsebuje ustrezno dokumentirane zvočne in video posnetke informativnih oddaj, ki jim je dodatno pridružen še jezikovni korpus besedil novic informativnih oddaj. Zbirka SiBN predstavlja tako poleg zbirke BNSI [Žgank-04] prvo takšno podatkovno zbirko v slovenskem jeziku.

V nadaljevanju bomo opisali samo lastnosti govornih podatkov, ki so vključeni v zbirki, jezikovni korpus pa bo predstavljen v naslednjem razdelku. Zbirka SiBN je še vedno v razvojni fazi. To pomeni, da se podatke zbirke še vedno dopolnjuje in izboljšuje transkripcije zvočnih posnetkov glede na različne namene uporabe. Trenutno zbirka vsebuje približno 34 ur ustrezno dokumentiranih posnetkov informativnih oddaj. Osnovno vodilo pri izbiri informativnih oddaj za označevanje v prvi fazi zbiranja podatkov je bilo, da bi bile oddaje čimbolj pestre po vsebini in čimbolj homogene po akustični kvaliteti. S tem smo hoteli pridobiti čimveč raznolikega govornega materiala, ki pa bi bil zajet v čimbolj konstantnih akustičnih pogojih, zato smo se v prvi fazi zbiranja odločili za posnetke samo ene informativne oddaje. Za oddajo smo izbrali osrednjo informativno oddajo nacionalne TV postaje RTVSLO1[6], *TV dnevnik*, ki se predvaja vsak dan ob 19:00 uri. Vsak TV dnevnik je sestavljen iz več vrst informativnih oddaj, ki vsebinsko pokrivajo različna področja dnevnih novic, v skupnem trajanju okoli ene ure. Trenutno je tako v zbirki SiBN obdelanih 34 TV dnevnikov, ki smo jih zajemali v času od maja do avgusta leta 2003.

Pridobivanje podatkov informativnih oddaj je obsegalo zajemanje zvočnih in video posnetkov ter delnih tekstovnih prepisov namenjenih podnaslavljanju TV dnevnikov. Zvočni posnetki so bili posneti enokanalsko pri frekvenci vzorčenja 16000 Hz in shranjeni v formatu WAV z uporabo 16–bitne linearne kvantizacije. Video posnetki so bili zajeti v standardnem formatu 25 slik na sekundo pri ločljivosti 320x240 točk na sliko. Shranjeni so v formatu *Windows Media Video* (WMV). Tekst namenjen podnaslavljanju informativnih oddaj smo pridobivali preko teleteksta. Zajemanje vseh treh podatkovnih tipov je potekalo hkrati s posebno strojno opremo in programskimi orodji,

[6] Z RTVSLO je bil sklenjen sporazum, ki je dovoljeval pridobivanje, označevanje in uporabo zvočnih posnetkov informativnih oddaj za raziskovalne namene.

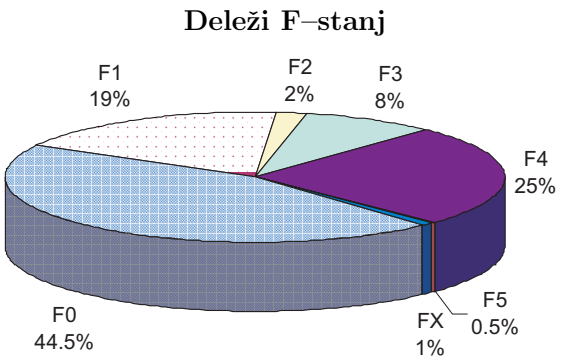
ki smo jih prilagodili za ta namen.

V nadaljevanju bomo analizirali zbirko glede na jezikovne in akustične lastnosti, ki so zajete v transkripcijah zvočnih posnetkov informativnih oddaj.

Tabela 2.3: Količina zvočnih posnetkov različnih tipov vsebin informativnih oddaj zbirke SiBN.

oddaja	skupno trajanje posnetkov	vsebina oddaje
Dnevnik I	10:59	novice iz Slovenije
	3:53	mednarodne novice
Dnevnik II	4:52	regionalne in lokalne novice
Denar	0:31	finančne in borzne informacije
Šport	3:18	športne novice
Vreme	1:38	vremenske napovedi
Magnet	1:43	kulturne novice
Skupaj	29:14	

Osnovna analiza vsebine informativnih oddaj je prikazana v tabeli 2.3. Skupna količina označenih govornih podatkov je nekaj več kot 29 ur, ostalih 5 ur posnetkov pripada neoznačenim blokom informativnih oddaj, kot so reklame, TV špice, odseki govora v tujem jeziku ipd. V označenem delu informativnih oddaj približno 90% govornih podatkov pripada novicam, ostalo pa pokrivajo najave vsebin informativnih oddaj ali najave novic. Med novicami največji delež zavzemajo novice iz Slovenije (približno 11 ur), pol manjši delež pripada regionalnim in lokalnim novicam, še nekaj manj pa mednarodnim novicam. Od ostalih vsebin zavzemajo največji delež športne informacije (12%), približno pol manjša deleža pa pripadata vremenskim napovedim (6%) in novicam iz sveta kulture (7%). Najmanjši delež označenih posnetkov predstavljajo borzne in finančne informacije. Vsega skupaj je bilo označenih 877 različnih novic in 148 sekcij z najavami. Strukturo vsebin, ki je prikazana v tabeli 2.3, smo uporabili tudi za osnovo pri hierarhični kategorizaciji novic za označevanje posameznih sekcij informativnih oddaj.



Slika 2.4: Deleži F–stanj v zbirki SiBN glede na skupno trajanje vsakega F–stanja.

Akustično raznolikost govornega materiala najbolje ponazarja razmerje med deleži F–stanj na sliki 2.4. Pretvorbo osnovnih transkripcij v F–stanja smo izvedli s postopkom,

ki je bil opisan v prejšnjem razdelku. Porazdelitev F–stanj v zbirki SiBN je pričakovana glede na to, da so v zbirko trenutno vključeni posnetki samo ene informativne oddaje. Največji delež govora pripada razredu F0 (44%), spontanega govora (F1) je okoli 19%, govora z različnimi akustičnimi ozadji (F4) pa okoli 25%. V primerjavi s podobnimi zbirkami informativnih oddaj v drugih jezikih [Pallett-02, Federico-00, Meinedo-01] lahko ugotovimo, da se razmerje med čistim govorom (F0) in spontanim govorom (F1) približno ujema z razmerji v teh zbirkah. Največja razlika pa je v deležu govora iz razreda F4. Običajno je delež takega govora enak deležu najbolj kvalitetnega govora iz razreda F0. V našem primeru pa imamo skoraj še enkrat več govora iz F0, kot je govora v razredu F4. Druga večja razlika je tudi v tem, da imamo sorazmerno malo telefonskih posnetkov (F2), in sicer samo 2%, kar pomeni približno pol ure govornega materiala. Večji delež pa zavzemajo posnetki z ozadjem glasbe (F3), kar je posledica dejstva, da imajo skoraj vse najave v TV dnevniku glasbo v ozadju. Največji delež F3 posnetkov prispevajo posnetki oddaj o kulturi, ki imajo skoraj vedno glasbeno spremljavo v ozadju. Posnetki iz razreda FX zavzemajo okoli 1% govornega materiala, kar ustreza podobnim zbirkam. Skoraj zanemarljiv pa je delež tujih govorcev, ki govorijo v jeziku transkripcij (F5), kar je pričakovano, saj slovenski jezik govori razmeroma malo tujcev.

Drugi pomemben pokazatelj akustične raznolikosti je porazdelitev in število različnih govorcev, ki so zajeti v govornem materialu. Statistika govorcev, ki so označeni v zbirki, je prikazana v tabelah 2.4 in 2.5 ter na sliki 2.5.

Tabela 2.4: Razporeditev govorcev po spolu v zbirki SiBN.

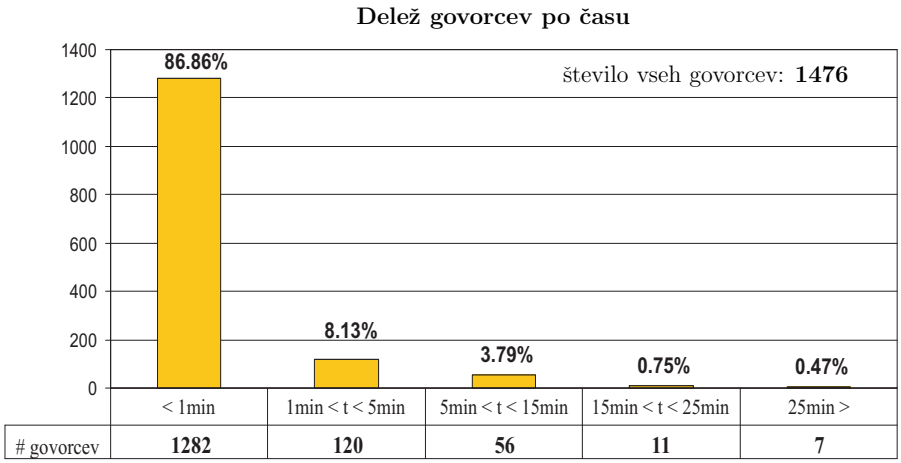
	govorci skupaj	moški govorci	ženske govorke
število govorcev	1476	1113	346
trajanje govora (hh:mm)	29:14	17:25	12:07

Tabela 2.5: Razporeditev govorcev glede na jezik v zbirki SiBN.

	govorci skupaj	materni jezik	ne-materni jezik	tuj jezik
število govorcev	1476	1165	50	261

Skupno število vseh govorcev, ki nastopajo v zbirki, je 1476. 1113 govorcev je moškega spola, ostalo pa so ženske govorke, ki prispevajo približno 41% govornega materiala. Razmerje v količini posnetkov med obema spoloma je tako primerljivo s podobnimi zbirkami v tujih jezikih [Graff-02], število govorcev pa je predvsem posledica slabše zastopanosti žensk v novicah informativnih oddaj. Razporeditev govorcev glede na jezik je prikazana v tabeli 2.5. Približno 80% govorcev pripada skupini, ki jim je slovenščina materni jezik, sledijo tujejezični govorci in tuji govorci, ki govorijo slovensko. Razmerje med govorci glede na jezik je pričakovano, saj večina novic, ki so označene v zbirki, pripada informacijam iz Slovenije ali pa lokalnim novicam. Razmerje med domačimi in tujimi govorci tako ustreza razmerju med domačimi in tujimi novicami.

Najboljši pokazatelj akustične raznolikosti po govorcih pa je prikazan na sliki 2.5, kjer je prikazana porazdelitev govorcev glede na skupno trajanje govora, ki ga je vsak

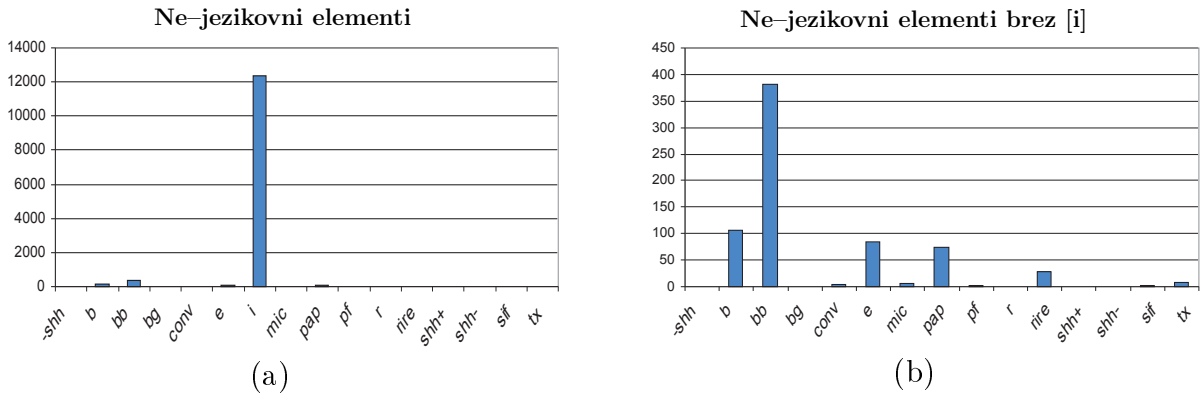


Slika 2.5: Porazdelitev govorcev glede na skupno trajanje njihovega govora v zbirki SiBN.

govorec prispeval v zbirki. Za zbirke informativnih oddaj je tipična ravno razporeditev govorcev po času. Največji delež govorcev pripada skupini, ki prispeva najmanj govora, po drugi strani pa imamo nekaj govorcev, ki prispevajo zelo velik delež govora. To so običajno voditelji informativnih oddaj. Tudi v našem primeru je tako. Zadnja dva stolpca v histogramu na sliki 2.5 pripadata voditeljem posameznih informativnih oddaj, pri katerih je vsak prispeval 15 minut govora ali več. Govorcev, ki so prispevali manj kot minuto govora, je približno 87%. V to skupino govorcev spadajo predvsem akterji novic, vse ostale skupine pa običajno vključujejo novinarje in napovedovalce poročil. To pomeni, da je govor govorcev iz prve skupine (prvi stolpec v histogramu) tudi najbolj raznolik in zajet v različnih akustičnih pogojih. To predstavlja tudi jedro večine težav pri nadaljnji obdelavi zvočnih posnetkov informativnih oddaj s postopki iz govornih tehnologij. Po eni strani imamo veliko različnih govorcev, ki jih moramo detektirati, prispevajo pa zelo malo govornega materiala in še to v različnih akustičnih razmerah. Zato je potrebno postopke prilagoditi tem razmeram, saj se pogosto izkaže, da postopki, ki delujejo v idealnih razmerah, ko imamo dovolj govornih podatkov v enakih akustičnih razmerah, povsem odpovejo v takšnih situacijah.

Ker je zbirka SiBN zasnovana predvsem za razvoj sistema za podnaslavljanje informativnih oddaj, smo analizirali tudi analizo besedišča in ne–govornih elementov v govornih prepisih zvočnih posnetkov.

Število besed, ki so zajete v tekstovnih prepisih govora, je 255 tisoč, od tega je 32 tisoč različnih besed. To pomeni precejšen skok v velikosti slovarja v primerjavi z dosedaj zbranimi govornimi zbirkami v slovenskem jeziku [Mihelič-03, Kačič-00]. Trenutno so izdelani slovarji fonetičnih prepisov približno 13 tisoč besed, ki se že uporabljajo pri izgradnji akustičnih modelov za razvoj sistema za samodejno prevajanje govora v okviru projekta *Voicetran* [Gros-05]. Poseben izziv sistemov za razpoznavanje govora pa zagotovo predstavlja spontani govor, ki je prav tako na voljo v zbirki SiBN. Pri tem je potrebno razpoznavalnike govora prilagoditi na številne posebnosti takega govora, predvsem na številne ne–jezikovne in ne–govorne elemente v govoru. Porazdelitev teh elementov v zbirki SiBN je prikazana na sliki 2.6.



Slika 2.6: Porazdelitev ne–jezikovnih elementov v posnetkih iz zbirke SiBN. Na sliki (a) je prikazana porazdelitev vseh ne–jezikovnih elementov, na sliki (b) pa porazdelitev brez elementov [i], ki označujejo dihanje govorcev.

Na sliki 2.6 (a) lahko vidimo, da je v zbirki označenih izrazito največ elementov, ki označujejo dihanje govorcev pri govorjenju. Pri tem moramo poudariti, da so bili premori, ko govorec zajame ali izdihne zrak, osnova za postavitev mej med osnovnimi segmenti govora v transkripcijah. Označevalci so bili še posebej pozorni na te elemente in so jih skrbno beležili. Na sliki 2.6 (b) je prikazana še porazdelitev ostalih ne–jezikovnih elementov. Tu največji delež pripada še drugemu najpogostejšemu pojavu pri govorjenju, to je tleskom ([bb]) in raznim drugim šumom ([b]), ki jih povzročamo z usti, ko govorimo. Ostali elementi nastanejo predvsem zaradi napak pri govoru oziroma so prisotni v ozadju govora. Pri tem moramo opozoriti, da tu niso zajeti medmeti in številna mašila, ki so prisotna predvsem pri spontanem govoru. Te smo posebej označevali neposredno v tekstovnih prepisih govora. Pri gradnji akustičnih modelov za razpoznavanje govora pa jih moramo posebno obravnavati, praviloma tako, da jih posebej modeliramo in pridružimo osnovnim modelom govora.

2.2.1 Jezikovni korpus zbirke SiBN

Kot smo že omenili, zbirka SiBN vsebuje tudi korpus besedil večjega števila informativnih oddaj. Namen korpusa je, da bi na podlagi besedil, ki smo jih pridobili neposredno iz različnih informativnih oddaj, pridobili jezikovne modele, ki bi jih lahko vključili v izgradnjo sistema za podnaslavljanje informativnih oddaj.

V ta namen smo zvočnim in video posnetkom informativnih oddaj pridružili še tekstovne prepise, ki so bili posredovani preko teleteksta in se uporabljajo za podnaslavljanje informativnih oddaj. Zajemanje besedil je potekalo v obdobju od maja 2003 do decembra 2004. Da bi bila besedila informativnih oddaj čimbolj usklajena z govornim materialom zbirke SiBN, smo se tudi tu odločili za pridobivanje tekstovnega materiala informativnih oddaj iste nacionalne TV postaje - RTVSLO1. Besedila, ki smo jih pridobivali, so bila besedila, namenjena podnaslavljanju informativnih oddaj, ki se vsakodnevno posredujejo preko teleteksta RTVSLO. Zajemanje besedil je potekalo sedemkrat na dan za vse informativne oddaje, ki so bile predvajane na TV postaji RTVSLO1.

Besedila predstavljajo delne podnapise različnih informativnih oddaj, od kratkih poročil do enournih sestavljenih informativnih oddaj. Namen zbiranja takšnih besedil je bil predvsem v tem, da bi dobili čim več tekstovnega materiala jezika, ki se uporablja pri TV informativnih oddajah. Posredovana besedila so tako predstavljala približne prepise govora, ki se uporabljajo pri informativnih oddajah, in je bila zato njihova kvaliteta izjemno slaba. Tekst je bil neenotno označen, besede so bile napačno zapisane, uporabljale so se številne okrajšave besed in besednih zvez, ki so bile neenotno izbrane, števila in rezultati so bili zapisani na različne načine, tuja lastna in zemljepisna imena so bila napačno zapisana ipd. Zato je bilo potrebno celoten zajeti tekstovni material ustrezno obdelati, popraviti besedišče in poenotiti oznake.

Trenutno smo tako obdelali 1358 prepisov informativnih oddaj v obsegu 280 dni iz obdobja od decembra 2003 do decembra 2004. Dodatno smo definirali tudi posebne jezikovne kategorije, ki smo jih označevali v tekstu, in sicer: kategorijo osnovnih števil, kategorijo vrstilnih številnikov, kategorijo rezultatov in drugih športnih izidov, kategorijo lastnih imen ter kategorijo fizikalnih enot in matematičnih količin. Skupno število vseh besed v korpusu tako znaša okoli 2 milijona 300 tisoč besed, kar pomeni okoli 110 tisoč različnih besed oziroma 68 tisoč besed ob uporabi kategorij.

Po številu vseh besed se tako naš jezikovni korpus uvršča med manjše korpuse besedil v slovenskem jeziku [JTSI-03], po številu različnih besed pa je povsem primerljiv z njimi. Pomembno je, da zbrani tekstovni podatki, ki smo jih pridobili neposredno iz prepisov govora informativnih oddaj, predstavljajo prvi tak korpus besedil v slovenskem jeziku in je primeren predvsem za izgradnjo jezikovnih modelov, ki se bodo uporabljali v sistemu za samodejno podnaslavljanje informativnih oddaj v slovenskem jeziku.

2.3 Večjezična zbirka informativnih oddaj COST278

Druga zbirka, ki smo jo uporabljali v našem raziskovalnem delu, je bila večjezična zbirka informativnih oddaj COST278. Namen in nastanek te zbirke je bil drugačen kot pri zbirki SiBN.

Zbirka COST278 [Vandecatseye-04] je nastala v okviru sodelovanja 10 raziskovalnih institucij v projektu COST278[7]. Pri zbirki je vsaka sodelujoča skupina prispevala po tri ure posnetkov informativnih oddaj. Tako je v zbirki zbranih in označenih približno 30 ur posnetkov informativnih oddaj v devetih evropskih jezikih: češkem, slovaškem, portugalskem, grškem, nizozemskem, hrvaškem, madžarskem, galicijskem in v slovenskem. Posnetki informativnih oddaj, ki prav tako kot v zbirki SiBN zajemajo zvočne in video posnetke, so zbrani iz štirinajstih različnih televizijskih postaj. Namen združevanja takšnih oddaj je bil, da bi pridobili čimbolj raznolike govorne in jezikovne vsebine, zajete v različnih akustičnih pogojih in v različnih časovnih obdobjih. Tako zbrane in enotno označene posnetke bi lahko uporabljali za razvoj postopkov govornih tehnologij neodvisnih od jezika obdelave, in za preučevanje njihove neobčutljivosti na različne pogoje delovanja.

[7]EU projekt COST Action 278: Spoken Language Interaction in Telecommunication, posebna skupina COST278 BN Interest Group.

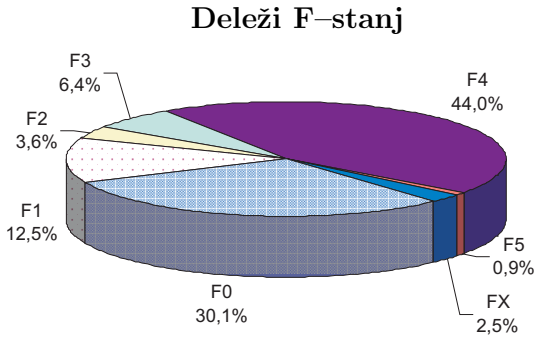
Bistveni poudarek združevanja posnetkov v zbirko je bil zato namenjen predvsem poenotenju transkripcij informativnih oddaj. V ta namen sta bila organizirana dva delovna srečanja, kjer je potekalo poenotenje transkripcij na več nivojih, in sicer v naslednjih elementih transkripcij:

- pri označevanju odsekov govorcev, kjer smo se držali pravil označevanja kvalitete in kanala posnetka iz tabele 2.2;
- pri tekstovnih prepisih govora, kjer smo sledili pravilom združenja LDC z dodatki, ki so bili že opisani v prejšnjih razdelkih;
- v postavljanju mej med osnovnimi segmenti: pravilo je bilo, da se postavi meja med segmentoma v točki vdiha; koliko mej se postavi, pa je bilo odvisno od dolžine premora med govorom;
- pri oznakah govorcev;
- pri označevanju odsekov najavnih in odjavnih TV špic in drugih multimedijskih dodatkov, ki spremljajo informativne oddaje;
- pri označevanju govora v tujem jeziku in
- pri označevanju ne–jezikovnih in ne–govornih elementov govora.

Pri vseh naštetih elementih označevanja smo v glavnem sledili pravilom združenja LDC, ki smo jih v splošnem že opisali v prejšnjih razdelkih.

Dodatno smo poenotili tudi formate zvočnih in video posnetkov. Podobno kot v zbirki SiBN je bila tudi tu frekvenca vzorčenja zvočnih signalov 16 kHz, snemanje je bilo enokanalsko, posnetki pa so shranjeni v WAV formatu. Video posnetki pa so bili zajeti v ločljivosti 352x288 in shranjeni v formatu *Real Media Video* (RM).

Ker so v zbirki zbrani podatki različnih informativnih oddaj v različnih jezikih, nismo izvajali vsebinske analize podatkov, ampak smo se osredotočili predvsem na analizo akustične raznolikosti posnetkov, ki jo bomo predstavili v nadaljevanju.



Slika 2.7: Deleži F–stanj v zbirki COST278 glede na skupno trajanje vsakega F–stanja.

Na sliki 2.7 so zbrani skupni deleži F–stanj vseh posnetkov informativnih oddaj v zbirki COST278. V primerjavi z zbirko SiBN tu lahko opazimo drugačna razmerja med F–stanji. Predvsem je opazna razlika v razmerju med govorom stanja F0 in govorom stanja F4. Delež govornih podatkov v obeh skupinah je ravno obraten, kot je bilo to v zbirki SiBN. Opazno sta se povečala tudi deleža govora iz F2 in FX, kar priča o tem, da so govorni podatki v tej zbirki zajeti v slabših akustičnih razmerah, kot so bili v primeru zbirke SiBN. To pa je bil tudi eden izmed ciljev nastanka te zbirke. Dodatno lahko opazimo, da je v primerjavi z zbirko SiBN upadla tudi količina spontanega govora in govora z ozadjem glasbe. V prvem primeru je to posledica dejstva, da je večina informativnih oddaj, ki so zbrane v zbirki, bolj osredotočena k mednarodnim novicam in manj k lokalnim poročilom. Glasba v ozadju govora pa je bolj specifična za zbirko SiBN, kjer imamo zbrane posnetke samo ene informativne oddaje, kjer je prisotno veliko glasbe v ozadju. Na splošno lahko ugotovimo, da razmerje deležev F–stanj ustreza zbirkam, kjer imamo zbrane posnetke informativnih oddaj v samo enem jeziku [Graff-02, Meinedo-03b, Federico-00].

Statistika govorcev v zbirki COST278 je podobna, kot je bila v zbirki SiBN in je zbrana v tabelah 2.6 in 2.7. Porazdelitev govorcev po času pa je prikazana v histogramu na sliki 2.8.

Tabela 2.6: Razporeditev govorcev po spolu v zbirki COST278.

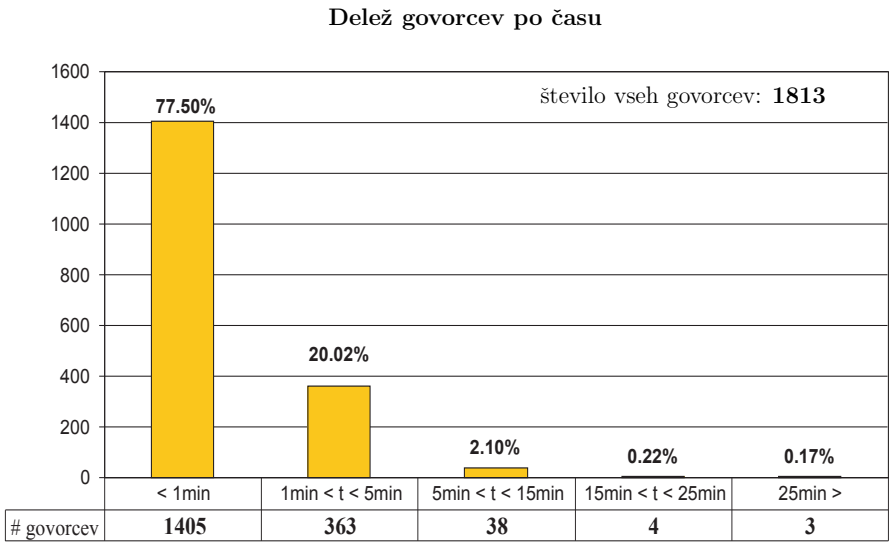
	govorci skupaj	moški govorci	ženske govorke
število govorcev	1815	1241	480
trajanje govora (hh:mm)	28:39	17:21	10:08

Tabela 2.7: Razporeditev govorcev glede na jezik v zbirki COST278.

	govorci skupaj	materni jezik	ne-materni jezik	tuj jezik
število govorcev	1815	1595	127	93

Razmerje med številom ženskih in moških govorcev (tabela 2.6) je podobno razmerju iz zbirke SiBN, delež ženskega govora pa se je tu nekoliko zmanjšal, in sicer na 35%. Nekoliko so se spremenila tudi razmerja med deleži govora v tujem in domačem jeziku (tabela 2.7). Delež govorcev v domačem jeziku je pričakovan in je primerljiv z zbirko SiBN. Zanimivo pa je, da se je v primerjavi s SiBN spremenilo razmerje med govorci v tujem jeziku in govorci, ki uporabljajo jezik transkripcije, ki pa ni njihov materni jezik.

Porazdelitev govorcev po času je prikazana na sliki 2.8 in povsem ustreza tipični porazdelitvi govorcev v zbirkah informativnih oddaj. Največ je govorcev, ki prispevajo najmanj govora, izredno malo pa imamo govorcev z veliko govora. Tako lahko ugotovimo, da je delež govorcev, ki prispevajo manj kot 5 minut govora, 97.5%, kar je za približno 2% več, kot je to v zbirki SiBN, čeprav je razmerje tistih, ki prispevajo več ali manj kot minuto govora manj izrazito, kot je to pri zbirki SiBN. To si lahko razlagamo z dejstvom, da imamo v zbirki COST278 zbrane različne informativne oddaje, zato je tudi skupnih govorcev zelo malo, kar posledično pomeni manjše deleže govorcev, ki



Slika 2.8: Porazdelitev govorcev glede na skupno trajanje njihovega govora v zbirki COST278.

prispevajo veliko govora. V tem primeru tudi ne moremo sklepati, kako se porazdeljujejo akterji novic in novinarji, ker imamo premalo skupnih posnetkov ene informativne oddaje, kot je bilo to v primeru zbirke SiBN.

Skupna ugotovitev je, da je zbirka COST278 veliko bolj raznolika v primerjavi z zbirko SiBN tako po vsebinskih, jezikovnih in akustičnih lastnostih, kar je bil tudi glavni namen zbirke. Bistveno v tej zbirki pa je bilo to, da smo poenotili transkripcije zvočnih posnetkov različnih informativnih oddaj.

2.4 Zaključek

V tem poglavju sta bili predstavljeni dve zbirki informativnih oddaj, ki smo jih uporabljali v okviru raziskovalnega dela doktorske disertacije.

Zbirka SiBN je zbirka informativnih oddaj v slovenskem jeziku in je bila zasnovana z namenom pridobivanja govornih in jezikovnih podatkov za razvoj sistema za samodejno podnaslavljanje informativnih oddaj, zato jo štejemo med govorne zbirke. Takšne zbirke predstavljajo drugačen koncept pridobivanja in označevanja zvočnih podatkov, kot je to v primeru vnaprej pripravljenih in načrtovanih govornih zbirkah. Zaradi velike količine podatkov in raznolikosti akustične, jezikovne in vsebinske informacije, je največji problem dokumentiranja takšnih zbirk ravno označevanje zvočnih posnetkov. V zbirki SiBN je trenutno popolnoma označenih 34 ur posnetkov ene informativne oddaje, poleg tega pa zbirka vključuje tudi obsežen jezikovni korpus, ki je sestavljen iz besedil, namenjenih podnaslavljanju informativnih oddaj.

Medtem ko je bilo potrebno zbirko SiBN pridobiti in označiti povsem na novo, pa je bilo v zbirki COST278 potrebno samo uskladiti transkripcije različnih informativnih oddaj v devetih evropskih jezikih. Zbirka COST278 je namreč nastala kot plod sodelo-

vanja več raziskovalnih institucij z namenom združevanja govornih podatkov različnih informativnih oddaj v različnih jezikih. Namen zbirke je bil, da bi poenotili in uskladili transkripcije zvočnih posnetkov informativnih oddaj, ki bi jih lahko uporabljali za razvoj postopkov obdelave govornih podatkov, ki so neodvisni od jezika.

V okviru raziskovalnega dela doktorske disertacije smo uporabljali obe zbirki za razvoj in testiranje postopkov segmentacije, detekcije govora in razvrščanja segmentov po govorcih. Razdelitev posnetkov na učne, razvojne in testne množice, ki smo jih uporabljali v posameznih preizkusih, je podrobneje opisana v dodatku A disertacije.

3 Detekcija govornih delov v zvočnih posnetkih

V tem poglavju bomo obravnavali razdelitev zvočnih posnetkov informativnih oddaj na govorne in ne–govorne dele. Naloga iskanja odsekov govora v zvočnih posnetkih je sestavljena iz dveh problemov: segmentacije zvočnih posnetkov in razvrščanja segmentov glede na govor in ne–govor. V tem poglavju se bomo ukvarjali s predstavitvami zvočnih posnetkov ter postopki segmentacije in razvrščanja segmentov na govorne in ne–govorne odseke na podlagi teh predstavitev. Tako bomo predstavili standardne postopke segmentacije in razvrščanja, ki temeljijo na akustičnih predstavitvah zvočnih signalov, in jih primerjali z novo predstavitvijo zvočnih signalov, ki je bila namensko razvita prav za ločevanje govornih in ne–govornih odsekov v zvočnih posnetkih. Glavna ideja predstavitve je bila v tem, da smo opazovali delovanje samodejnih razpoznavalnikov glasov na govornih in ne–govornih posnetkih in na podlagi glasovnih transkripcij pridobljenih iz razpoznavalnikov glasov izpeljali štiri osnovne fonetične značilke, primerne za določevanje govora in ne–govora.

V tem poglavju bomo predstavili dva tipa fonetičnih predstavitev, ki smo jih tvorili iz osnovnih zaporedij razpoznanih govornih enot iz dveh skupin razredov glasov. Značilke so bile izvedene tako, da so bile primerne za uporabo v sistemih za detekcijo govornih odsekov v zvočnih posnetkih. Primerjava standardnih - akustičnih in predlaganih - fonetičnih predstavitev zvočnih posnetkov je bila izpeljana v primeru dveh različnih postopkov detekcije govora, s katerimi se bomo prav tako ukvarjali v tem poglavju. Dodatno smo eksperimentirali tudi z združevanjem različnih predstavitev zvočnih posnetkov in v primeru združevanja akustičnih in predlaganih fonetičnih značilk dosegli najboljše rezultate ločevanja govornih in ne–govornih odsekov na zbirkah SiBN in COST278.

3.1 Uvod

Osnovna naloga segmentacije zvočnih posnetkov na govorne in ne–govorne odseke (*ang. speech/non–speech segmentation*) je razdeliti zvočne posnetke na dele, kjer je govor in na dele, kjer govora ni. Medtem ko so odseki govora dobro definirani - to so namreč področja v zvočnih posnetkih, kjer je prisoten govor enega ali več govorcev - zajemajo odseki ne–govora vse ostale dele zvočnih posnetkov, kjer govora ni, in so tako lahko sestavljeni iz različnih akustičnih pojavov, kot so npr. glasba, tišina, različni zvoki strojev in živali, šumi v ozadju ipd.

Običajno nas pri obdelavi zvočnih posnetkov v govornih aplikacijah zanimajo le govorni deli, zato lahko v tem primeru govorimo o detekciji govora v zvočnih posnetkih. Detekcija govora je primerna povsod, kjer nas zanima samo informacija iz govornih signalov in se hočemo znebiti nepotrebne obdelave ne–govornih delov. Z uporabo zanesljivih postopkov iskanja in detekcije govora tako zagotovimo razdelitev zvočnih posnetkov na dele, ki jih obdelujemo, in na dele, ki jih zavržemo. Na ta način zagotovimo bolj učinkovito delovanje sistemov za obdelavo govora in hkrati strukturiramo daljše zvočne posnetke na smiselne in razumljive odseke. Postopki detekcije in iskanja govora v zvočnih posnetkih se tako uporabljajo na različnih področjih uporabe govornih tehnologij: v splošnih sistemih za samodejno razpoznavanje govora [Shafran-03], v sistemih za samodejno podnaslavljanje informativnih oddaj [Gauvain-02, Woodland-02, Beyerlein-02], v sistemih indeksacije zvočnih posnetkov in samodejne izdelave povzetkov novic [Makhoul-00, Magrin-Chagnolleau-02], v sistemih iskanja in sledenja govorcev [Reynolds-05, Zhu-05, Sinha-05, Istrate-05, Moraru-05] ipd. Zanesljivo iskanje govora v daljših zvočnih posnetkih nam zagotavlja dobro delovanje postopkov obdelave govora v nadaljevanju, hkrati pa nam zmanjša čas obdelave zvočnih posnetkov. Zaradi tega je poleg zanesljivega delovanja postopkov detekcije govora potrebno zagotoviti tudi učinkovito in robustno delovanje teh postopkov, ki morajo biti izvedeni tako, da jih lahko brez težav vključimo v različne sisteme za obdelavo govora. Z drugimi besedami to pomeni, da moramo načrtovati takšne postopke detekcije govora, ki delujejo hitro, v različnih akustičnih pogojih in jih lahko vgradimo v različne sisteme govornih tehnologij.

V nadaljevanju bomo tako predstavili postopke detekcije in iskanja govora v zvočnih posnetkih informativnih oddaj, ki ustrezajo vsem zgoraj naštetim lastnostim. Pri tem se bomo osredotočili na predstavitve zvočnih signalov, ki so primerne za ločevanje govora in ne–govora ter na postopke iskanja oziroma segmentacije zvočnih posnetkov na podlagi teh predstavitev na govorne in ne–govorne odseke. Iskanje govora v zvočnih posnetkih združuje dva problema: segmentacijo in razvrščanje segmentov na govor in ne–govor. Dosedanje raziskovalno delo na tem področju je bilo omejeno predvsem na razvoj postopkov in predstavitev signalov samo za razvrščanje posnetkov na govor in ne–govor. Pri tem se je predpostavljalo, da so zvočni posnetki že segmentirani in je bilo potrebno v danih segmentih poiskati tiste, ki predstavljajo govor.

Dosedanje predstavitve signalov za razvrščanje segmentov na govorne in ne–govorne dele v glavnem temeljijo na akustičnih lastnostih segmentov. Pri tem se za predstavitve signalov uporabljajo standardni postopki izpeljave akustičnih značilk, ki jih uporabljamo tudi pri drugih postopkih obdelave govora. Za najbolj učinkovite so se

izkazale značilke koeficientov melodičnega kepstra MFCC, ki se uporabljajo predvsem za razpoznavanje govora [Picone-93]. Uporaba akustičnih predstavitev signalov za razvrščanje govora in ne–govora je predvsem posledica tradicije modeliranja govornih signalov za namene razpoznavanja govora, sloni pa na predpostavki, da lahko preko akustičnih predstavitev signalov modeliramo akustične vire, ki proizvajajo te signale. Pri govoru tako lahko modeliramo en vir - vir govora, problem pa nastane pri modeliranju ne–govora, kjer je virov lahko več. Običajno se zato pri razvrščanju segmentov na govor in ne–govor uporablja več modelov za detekcijo ne–govora in enega ali več za detekcijo govora. Drugačen pristop k modeliranju govora je obravnava govora kot zaporedja razpoznanih govornih enot. Tako lahko proces tvorjenja govora modeliramo kot končni avtomat s stanji, ki predstavljajo osnovne govorne enote [Ajmera-03]. Takšen avtomat bo drugače deloval, če bomo z njim opisovali govorne oziroma ne–govorne signale. Na podlagi karakterističnih lastnosti delovanja avtomata v primeru govornih in negovornih posnetkov tako lahko pridobimo značilke, primerne za ločevanje govora od ne–govora. Takemu pristopu modeliranja smo sledili tudi mi in na podlagi tega izvedli novo predstavitev signalov primerno za ločevanje govornih in ne–govornih delov, ki je temeljila na samodejno pridobljenih fonetičnih transkripcijah signalov. V nadaljevanju bomo predstavili predlagani postopek izpeljave fonetičnih značilk, ki so bile izpeljane iz razpoznanih osnovnih enot govora. Podali bomo osnovni koncept pridobivanja takšnih značilk in predstavili dve skupini značilk. Prva je bila izpeljana na podlagi zaporedja parov glasovnih enot samoglasnik–soglasnik, druga pa iz parov zvenečih in nezvenečih glasovnih enot.

Drugi del tega poglavja bo posvečen predvsem segmentaciji zvočnih posnetkov na govorne in ne–govorne dele *(segmentacija govor/ne–govor, GNG segmentacija)*. Za segmentacijo smo uporabljali dva postopka. Prvi postopek je bil povzet po [Ajmera-03] in smo ga ustrezno prilagodili za izvajanje segmentacije na podlagi naših predstavitev. Postopek je temeljil na izgradnji GMM modelov, ki jih v mreži HMM modelov uporabimo za razvrščanje in segmentacijo hkrati. Drugi postopek pa smo razvili prav zaradi predlaganih predstavitev zvočnih signalov. Izkazalo se je namreč, da predstavitve delujejo bolje, če so ocenjene iz daljših odsekov zvočnih posnetkov, zato smo segmentacijo in razvrščanje ločili na dva dela. V prvi fazi smo izvajali segmentacijo posnetkov glede na različne akustične lastnosti. S tem smo pridobili večje segmente zvočnih posnetkov, na katerih smo ocenili predlagane fonetične značilke, in na podlagi teh predstavitev smo v drugi fazi razvrstili segmente na govor in ne–govor. Tudi v tem primeru smo za razvrščanje segmentov uporabljali GMM modele. Postopki segmentacije bodo predstavljeni v razdelku 3.3.

V zadnjem delu se bomo posvetili vrednotenju različnih predstavitev in postopkov GNG segmentacije na zbirkah zvočnih posnetkov informativnih oddaj SiBN in COST278. Primerjali smo dve referenčni metodi z dvema različnima predstavitvama signalov z našimi predlaganimi postopki, ki so temeljili na fonetičnih predstavitvah. Dodatno pa smo izvedli še poskuse združevanja različnih predstavitev zvočnih posnetkov, s katerimi smo dosegli najboljše rezultate razpoznavanja govornih in ne–govornih segmentov, in jih bomo prav tako predstavili v nadaljevanju.

3.2 Pridobivanje značilk za detekcijo govora v zvočnih posnetkih

V tem razdelku bomo predstavili nov način pridobivanja značilk za razpoznavanje govornih in ne–govornih odsekov zvočnih posnetkov. Predstavitve signalov s temi značilkami bomo imenovali fonetične predstavitve in značilke, fonetične značilke. Predstavili bomo štiri mere za izračun značilk na podlagi samodejno pridobljenih osnovnih glasovnih enot govora, ki smo jih pridobivali neposredno iz razpoznavalnikov glasov. Na ta način bomo izpeljali dve skupini značilk: prva bo temeljila na kombinaciji parov samoglasnikov in soglasnikov, druga pa na parih zvenečih in nezvenečih glasov.

V nadaljevanju bomo najprej predstavili osnovne koncepte in motivacije za izvedbo takšnih značilk, nato pa opisali samo izpeljavo značilk.

3.2.1 Osnovni koncepti pridobivanja značilk

Osnovna ideja pri izvedbi značilk za razpoznavanje govornih in ne–govornih odsekov v zvočnih posnetkih je bila, da bi pridobili takšne predstavitve signalov, s katerimi bi dobro detektirali govor, hkrati pa bi približno enako opisovali različne ne–govorne pojave. S takšnimi predstavitvami bi tako pridobili značilke, ki bi bile primerne za modeliranje samo dveh osnovnih razredov razvrščanja, razreda govora in razreda ne–govora.

Običajno so namreč osnovni sistemi za razpoznavanje govora in ne–govora sestavljeni iz več statističnih modelov, s katerimi modeliramo različne akustične pojave v signalih, od govora do glasbe, tišine, različnih šumov ipd. Detekcija govornih odsekov v zvočnih posnetkih tako poteka na podlagi odločanja med modeli, ki jih predhodno ocenimo s pomočjo učnih podatkov. Ker so v takšnih sistemih običajno zvočni posnetki predstavljeni samo z akustičnimi predstavitvami, moramo tako pridobiti in ustrezno modelirati različne govorne in ne–govorne pojave in situacije. To pa predstavlja glavni problem takšnih postopkov. Vedno se je namreč potrebno odločati na podlagi več modelov, hkrati pa je potrebno predvideti vse možne situacije akustičnih pojavov, kar je v praksi skoraj nemogoče. Po drugi strani pa nas pri detekciji govora zanimajo le govorni odseki v zvočnih posnetkih in ne različni ne–govorni pojavi, zato lahko problem detekcije govora predstavimo kot problem odločanja, ali dani odsek predstavlja govor ali ne. V tem primeru gre torej za razvrščanje odsekov v dva razreda, pri katerem prvi razred - govor - definira drugega - ne–govor. Tako so nas zanimale takšne predstavitve zvočnih posnetkov, s katerimi bi dobro modelirali vsak razred samo z enim modelom.

Prvi poskus v tej smeri je izvedel Greenberg [Greenberg-95] z značilkami, ki so temeljile na opisovanju tipičnih spektralnih potekov signalov govora glede na pričakovan ritem sprememb zlogov v govoru. Karnebäck [Karnebäck-02] je prav tako uporabljal različne značilke, ki so bile izpeljane iz ocen frekvenc modulacije spektra govora, in je ugotovil, da v kombinaciji z MFCC značilkami predstavljajo robustne predstavitve za detekcijo govora in glasbe. Povsem drugačen pristop izpeljave značilk sta izvedla Williams in Ellis [Williams-99]. Zgradila sta samodejni razpoznavalnik glasov in opazovala njegovo

delovanje v primeru govora in glasbe. Na podlagi razlik v delovanju sta izpeljala značilke, ki so temeljile na ocenjevanju verjetnosti stanj razpoznavalnika glede na položaj in tip signala. Tako sta predlagala dve osnovni značilki, entropijo *(ang. entropy)* in dinamizem *(ang. dynamism)*, ki sta ju uporabljala za razvrščanje posnetkov na govor in glasbo. Ajmera s sod. [Ajmera-03] pa je na podlagi teh značilk izvedel sistem za segmentacijo zvočnih posnetkov na govor in glasbo, ki smo ga uporabljali tudi v okviru doktorske disertacije za GNG segmentacijo in ga bomo predstavili v nadaljevanju.

Idejo, da se pri detekciji govora uporabljajo značilke izpeljane na podlagi različnega delovanja razpoznavalnikov govora v primeru govornih in ne–govornih posnetkov, smo pri predlaganih fonetičnih predstavitvah razvili še naprej [Žibert-06]. Značilke smo namreč izpeljali na podlagi samodejnih transkripcij zvočnih posnetkov, ki smo jih pridobili neposredno iz razpoznavalnikov glasov. Osnovno vodilo pri tem je bilo, da je govor sestavljen iz smiselnih zaporedij osnovnih govornih enot, medtem ko so v primeru ne–govora zaporedja razpoznanih govornih enot bolj naključna.

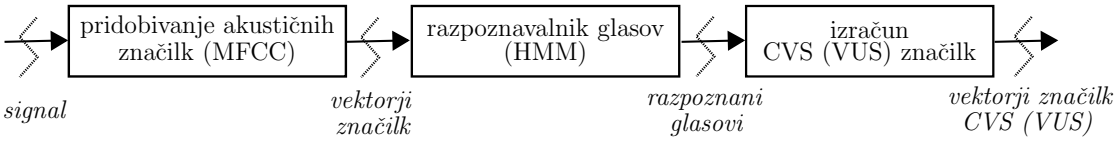
3.2.2 Fonetične značilke za detekcijo govora

V splošnem lahko razpoznavalnik glasov (govora) predstavimo kot sistem, ki zvočnim signalom na vhodu pripisuje zaporedja osnovnih govornih enot na izhodu. Vhodni signali so običajno parametrizirani z akustičnimi predstavitvami, ki so sestavljene iz zaporedij vektorjev značilk. V procesu razpoznavanja se tako zaporedje vektorjev značilk prevede v najbolj verjetno zaporedje osnovnih govornih enot, ki so vključene v razpoznavalnik. Poleg zaporedja osnovnih govornih enot pa je izhod razpoznavalnika glasov običajno opremljen tudi z informacijo o trajanju in verjetnostjo posamezne enote v zaporedju. Zato lahko tudi informacije, ki jih pridobimo s takšnimi razpoznavalniki, obravnavamo kot predstavitve vhodnih signalov. Seveda so razpoznavalniki glasov (govora) namenjeni predvsem obdelavi govornih signalov, lahko pa jih prav tako uporabljamo tudi na ne–govornih signalih. V primeru govornih signalov tako lahko pričakujemo pričakovano delovanje razpoznavalnikov in s tem tudi smiselne izhodne informacije, v primeru ne–govornih signalov pa je delovanje nepredvidljivo, kar se odraža tudi pri razpoznanih govornih enotah na izhodu. In ravno to dejstvo smo želeli izkoristiti pri izpeljavi fonetičnih značilk.

Po drugi strani pa je izhodna informacija, ki jo pridobimo iz transkripcij[1], močno odvisna od jezika govora, ki ga razpoznavamo, in modelov, ki so vključeni v razpoznavalnik. Dodatno omejitev predstavlja tudi dejstvo, da v splošnem razpoznavalniki glasov ne delujejo zelo zanesljivo v različnih akustičnih razmerah. To pomeni, da je bilo potrebno načrtovati takšne značilke, ki bi bile neodvisne od jezika govora in zanesljivosti delovanja razpoznavalnikov, zato smo se odločili, da bomo opazovali delovanje razpoznavalnikov iz transkripcij na podlagi širših skupin glasov. Izvedli smo dve skupini fonetičnih značilk, prva je temeljila na opazovanju zaporedja parov glasovnih enot

[1] Transkripcije v primeru fonetičnih značilk za detekcijo govora zajemajo razpoznane glasovne enote, čas trajanja posameznih enot in oceno verjetnosti posamezne enote (opcijsko) v razpoznanem zaporedju enot, kar bo podrobneje opisano v nadaljevanju. Zato tudi tu, podobno kot v primeru transkripcij zvočnih posnetkov informativnih oddaj, uporabljamo besedo transkripcija namesto glasovnih prepisov.

soglasnik–samoglasnik *(ang. consonant–vowel, CV)*, druga pa je bila izpeljana iz parov zvenečih in nezvenečih glasovnih enot *(ang. voiced–unvoiced, VU)*. Izbira širših skupin glasov je smiselna vsaj iz dveh razlogov. Ker so te skupine glasov značilne za večino jezikov, so tudi izpeljane značilke manj odvisne od jezika govora, hkrati pa na ta način povečamo zanesljivost ocen značilk, saj je razpoznavanje širših skupin glasov običajno boljše kot razpoznavanje osnovnih govornih enot. Drugi razlog je bil ta, da lahko govor obravnavamo tudi kot zaporedje značilnih zlogov sestavljenih iz osnovnih govornih enot, iz katerih tvorimo besede. To pomeni, da ocenjevanje delovanja razpoznavalnika opazujemo na nivoju zlogov in ne na nivoju osnovnih govornih enot, s čimer še dodatno povečamo neobčutljivost značilk na napake razpoznavanja.



Slika 3.1: Shema pridobivanja CVS (VUS) značilk za detekcijo govora v zvočnih posnetkih.

Shema pridobivanja fonetičnih značilk je prikazana na sliki 3.1. Kot smo že omenili, signal najprej parametriziramo z akustičnimi značilkami, ki so primerne za razpoznavanje govora. Običajno se za razpoznavanje govora uporabljajo osnovne značilke MFCC z izpeljanimi prvimi in drugimi odvodi [Picone-93]. Razpoznavalnik glasov nato na podlagi zaporedja vektorjev značilk tvori transkripcijo signala, ki je sestavljena iz najbolj verjetnega zaporedja osnovnih govornih enot s pridruženim trajanjem posameznih enot v zaporedju. V našem primeru smo pri vseh izpeljavah fonetičnih značilk uporabljali razpoznavalnike glasov, ki so temeljili na HMM modelih. Transkripcije se nato prevedejo v širše skupine glasov, ki jih predhodno določimo glede na tip razpoznavalnika. V primeru parov soglasnik–samoglasnik imamo tri skupine glasov: samoglasnike (V), soglasnike (C) in oznake za premore (S). Te enote označujemo kot CVS enote. V primeru zvenečih in nezvenečih glasov pa dobimo naslednje skupine glasov: zveneče (V), nezveneče (U) in premore (S). Te enote označujemo kot VUS enote. Po preslikavi osnovnih govornih enot v CVS (VUS) enote se izvaja izpeljava fonetičnih značilk. V primeru CVS enot govorimo o CVS značilkah, v primeru VUS enot pa o VUS značilkah. V tej fazi postopka se izvaja analiza transkripcij na podlagi CVS ali VUS enot in izračun fonetičnih značilk, ki jih po izračunu sestavimo v vektorje CVS (VUS) značilk.

Pri izvedbi CVS (VUS) značilk smo iskali tiste lastnosti CVS (VUS) transkripcij, s katerimi je bilo možno ločevati govorne in ne–govorne posnetke. Dodatno smo pazili tudi na to, da bi bile značilke čim manj odvisne od napak razpoznavanja. Pri tem smo izvajali analizo transkripcij na podlagi več razpoznavalnikov glasov, namenjenih razpoznavanju govora v različnih jezikih. Po natančni analizi delovanja razpoznavalnikov smo tako določili štiri mere za izpeljavo značilk, ki so temeljile na *trajanju* in *spremembah* posameznih enot v transkripcijah [Žibert-06]. Te so:

- *Normirano razmerje trajanja CV (VU) enot*, ki ga izračunamo kot:

$$\frac{|t_C - t_V|}{t_{CVS}} + \alpha \cdot \frac{t_S}{t_{CVS}}, \tag{3.1}$$

kjer predstavlja t_C skupno trajanje vseh razpoznanih soglasnikov (C) v oknu analize s trajanjem t_{CVS}, t_V pa skupno trajanje vseh samoglasnikov (V). Z drugim členom v izrazu (3.1) merimo s trajanjem t_S prisotnost premorov (S) v obravnavanem signalu. Z utežjo α uravnavamo razmerje med prisotnostjo razpoznanih govornih enot in detektiranih premorov v signalu. Utež α izbiramo iz intervala $[0, 1]$. Običajno jo postavimo kar na 0.5, sicer pa jo lahko določimo glede na optimalne rezultate segmentacije. V primeru VUS enot soglasnike zamenjamo z nezvenečimi glasovi (U), samoglasnike z zvenečimi (V), premori (S) pa ostanejo enaki.

Pri izpeljavi te mere smo upoštevali dejstvo, da je govor sestavljen iz kombinacije CV (VU) enot, ki se prepletajo z relativno kratkimi premori (S). Zato lahko pričakujemo v primeru govornih signalov približno enaka skupna trajanja CV (VU) enot in kratko skupno trajanje premorov. To pa pomeni, da je v primeru govora vrednost izraza (3.1) okoli 0.0. Ravno nasprotno pa je v primeru ne–govornih signalov. Ker se CV (VU) enote v transkripcijah ne–govora skoraj nikoli ne porazdeljujejo enakomerno, je tudi razlika v skupnem trajanju posameznih enot večja. Prav tako se tudi večina daljših premorov v signalih razpozna kot premor, zato je tudi relativno trajanje premorov večje. Iz tega sledi, da se v primeru ne–govornih signalov vrednost izraza (3.1) giba okoli 1.0.

Tu moramo omeniti, da smo pri računanju razmerji med CV (VU) enotami v izrazu (3.1) uporabili absolutne razlike med trajanji ($|t_C - t_V|$) in ne pravih deležev podanih z razmerji t_C/t_V ali t_V/t_C. V primeru deležev bi namreč vedno ena enota (C ali V) prevladovala nad drugo, kar bi v primeru različnih razpoznavalnikov pomenilo različne vrednosti izraza (3.1) in s tem bi dobili slabše ocene predlagane mere.

- *Normirana CV (VU) hitrost govora* je definirana z izrazom:

$$\frac{n_C + n_V}{t_{CVS}}, \tag{3.2}$$

kjer sta n_C in n_V števili razpoznanih C in V enot v signalu v času t_{CVS}. Pri tem ne upoštevamo število razpoznanih premorov (enot S) v signalu. V primeru VU enot je mera (3.2) definirana podobno.

S to mero ocenjujemo hitrost govora na nivoju osnovnih govornih enot. Dejstvo je, da, ko govorimo, tvorimo zaporedje osnovnih govornih enot, ki se spreminjajo na vsakih nekaj deset milisekund. S štetjem teh sprememb na nekem časovnem intervalu tako dobimo hitrost govora, ki je odvisna predvsem od posameznega govorca in tipa govora, zato se takšne značilke uporabljajo tudi v sistemih za razpoznavanje govorcev [Reynolds-03b]. V našem primeru smo oceno hitrosti govora uporabljali za ločevanje med govornimi in ne–govornimi signali. Izkazalo se je namreč, da kljub temu, da se hitrost govora spreminja glede na različne govorce in tipe govora, se še vedno spreminja drugače kot v primeru ne–govora. V primeru ne–govornih posnetkov smo namreč opazili, da je spreminjanje osnovnih govornih enot veliko manjše kot v primeru govora. Pri tem moramo poudariti, da nismo šteli sprememb enot S. S tem smo se hoteli znebiti vpliva spontanega govora, kjer je hitrost govora tudi zaradi številnih premorov nižja.

Mera hitrosti govora, definirana z izrazom (3.2), se obnaša podobno kot mera povprečnega dinamizma, ki je bila predlagana v [Williams-99] za ločevanje govora in glasbe.

- *Normirane spremembe CVS (VUS) enot* so definirane z izrazom

$$\frac{c(C,V,S)}{t_{CVS}}, \tag{3.3}$$

kjer v funkciji $c(C,V,S)$ štejemo, kolikokrat je prišlo do zamenjave ene izmed enot C, V, S z drugo v času t_{CVS}. Podobno štejemo spremembe v primeru VUS enot s funkcijo $c(V,U,S)$.

S to mero podobno kot v prejšnjem primeru merimo spremembe CV (VU) enot v signalu, vendar obstaja pomembna razlika med obema merama. V prejšnjem primeru smo šteli vse spremembe med enotami v signalu, tu pa štejemo samo spremembe med različnimi enotami. Na ta način v bistvu ocenjujemo, koliko zlogov CV (VU) je prisotnih na nekem odseku signala. Dejstvo je namreč, da je govor sestavljen iz osnovnih zlogov CV (VU), zato lahko v tem primeru pričakujemo večje vrednosti ocene izraza (3.3). Po drugi strani pa smo iz analiziranih transkripcij ne–govornih signalov ugotovili, da ne vsebujejo veliko takšnih zlogov, kar posledično pomeni nižje vrednosti mere (3.3).

Predlagano mero lahko razširimo še naprej. V našem primeru smo opazovali samo pare enot C (U) ,V (V), S (S), lahko pa bi šteli tudi višje kombinacije enot. V tem primeru bi tako dobili *n–gramske* modele CVS (VUS) enot (kot pri izgradnji statističnih jezikovnih modelov), ki bi jih lahko ocenjevali iz govornih in ne–govornih transkripcij.

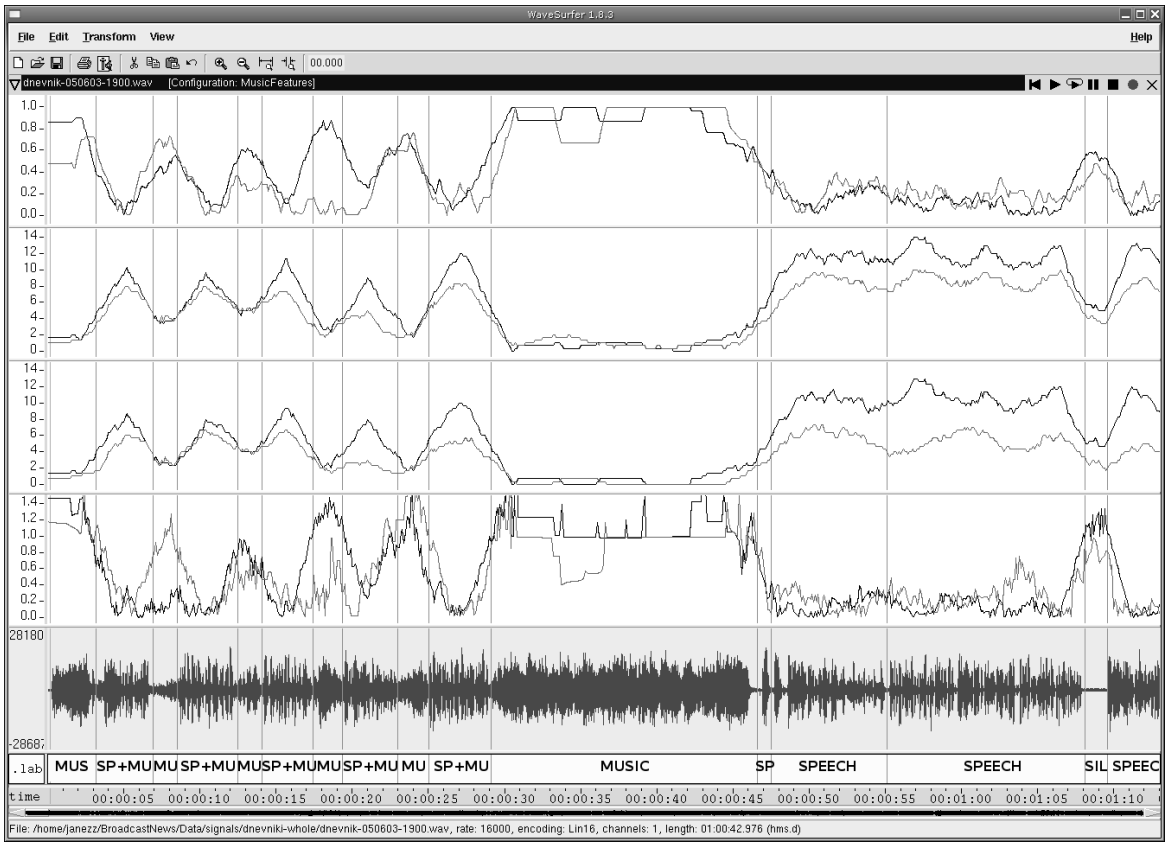
- *Normirana razlika povprečnega trajanja CV (VU) enot* je definirana z izrazom

$$\frac{|\bar{t}_C - \bar{t}_V|}{\bar{t}_{CV}}, \tag{3.4}$$

kjer s $\bar{t}_C$ in $\bar{t}_V$ označujemo povprečno trajanje C in V enot na danem odseku analiziranega signala. Podobno tudi $\bar{t}_{CV}$ predstavlja povprečje trajanja enot (C,V) v danem odseku signala. V primeru VU enot namesto soglasnikov (C) uporabljamo nezveneče glasove (U) in namesto samoglasnikov (V) zveneče glasove (V).

S to značilko merimo razmerje med povprečnim trajanjem samoglasnikov (zvenečih glasov) in soglasnikov (nezvenečih glasov). Znano je, da so samoglasniki pri govoru v povprečju daljši od soglasnikov, podobno je tudi v primeru zvenečih in nezvenečih glasov, zato lahko v primeru govornih signalov pričakujemo značilna razmerja med temi enotami. V primeru ne–govornih signalov pa nismo opazili takšnih lastnosti, saj so bila razmerja v povprečnem trajanju med CV (VU) enotami precej raznolika.

Mera v (3.4) je korelirana z mero (3.1), saj v obeh primerih primerjamo razmerje med trajanji CV (VU) enot. Razlika pa je v tem, da v tem primeru vzamemo povprečno trajanje enot, v prejšnjem pa smo vzeli skupno trajanje enot. Tudi v tem primeru smo uporabljali razliko med trajanji in ne deležev. Razlogi za to so podobni kot v prejšnjem primeru.



Slika 3.2: Potek CVS značilk. Zgornje/prvo okno prikazuje značilko normiranega razmerja trajanja CV enot, drugo okno prikazuje normirano CV hitrost govora, tretje normirane spremembe CVS enot, v četrtem oknu pa je prikazan potek značilke normirane razlike povprečnega trajanja CV enot. V vsakem oknu sta prikazana dva poteka: temnejša črta predstavlja delovanje značilk, ki smo jih pridobili iz slovenskega razpoznavalnika glasov, svetlejša črta pa prikazuje potek značilk ob uporabi angleškega razpoznavalnika glasov. V spodnjem oknu je prikazan zvočni signal skupaj z oznakami govornih in ne–govornih delov.

Z vsemi štirimi predlaganimi značilkami merimo posamezne lastnosti transkripcij na podlagi CVS (VUS) enot na določenih odsekih signalov, ki jih obdelujemo. Ti odseki morajo biti dovolj veliki, da lahko dovolj dobro ocenimo vrednosti značilk. V naših preizkusih smo uporabljali vnaprej določene odseke dolžin od 2.0 do 5.0 s. Lahko pa definiramo odseke tudi s številom razpoznanih enot, na katerih ocenjujemo predlagane značilke. Izbira odsekov je seveda odvisna od namena uporabe in od pričakovane dolžine govornih in ne–govornih odsekov v zvočnih posnetkih. V našem primeru smo se morali omejiti na krajše odseke, dodatna zahteva pa je bila, da smo hoteli pridobivati CVS (VUS) značilke, ki bi bile časovno usklajene z akustičnimi značilkami. To pa zato, ker smo hoteli združevati oba tipa predstavitev v postopkih fuzije GNG segmentacije. Časovno usklajevanje fonetičnih značilk z akustičnim značilkami smo dosegli na podoben način, kot to delamo v primeru spektralne analize signalov. Tako smo poleg osnovnega odseka (okna) izračuna značilk definirali še premik po času za naslednji izračun *(ang. frame skip)*. S tem smo tako računali značilke na konstantnih oknih analize na vsakih nekaj milisekund (odvisno od premika). Na drugačen način pa smo računali

fonetične značilke, ki smo jih pridobili iz vnaprej segmentiranih zvočnih posnetkov. V tem primeru ni bilo potrebno definirati oken in premika izračunov, ampak smo ocenili CVS (VUS) značilke kar na segmentu posnetka. Tako pridobljene značilke seveda niso bile časovno usklajene z akustičnimi značilkami.

Na sliki 3.2 je prikazan potek CVS značilk v primeru krajšega zvočnega posnetka slovenske informativne oddaje iz zbirke SiBN[2]. Potek CVS značilk je bil izveden ob uporabi dveh različnih razpoznavalnikov glasov: prvi je bil slovenski razpoznavalnik glasov (temnejša črta na potekih značilk), drugi pa je bil naučen na zbirki TIMIT [Garofolo-93] in je bil namenjen razpoznavanju angleških govornih enot (svetlejša črta). Angleški razpoznavalnik glasov je bil v tem primeru uporabljen na slovenskem govoru. Kot lahko vidimo iz slike 3.2, je zvočni posnetek sestavljen iz različnih odsekov govora in ne–govora. Govorni odseki vključujejo govor z ozadjem glasbe (SP+MU) in čist govor različnih govorcev (SPEECH), ne–govorni odseki pa so sestavljeni iz glasbe (MUSIC) in tišine (SIL). V zgornjih oknih na sliki 3.2 so prikazani poteki posameznih značilk iz (3.1) - (3.4) v primeru CVS enot. Iz potekov značilk na sliki 3.2 so razvidna velika odstopanja značilk v primeru govornih in ne–govornih odsekov. Dodatno lahko ugotovimo tudi, da je potek približno enak za različne govorne in ne–govorne odseke. To potrjuje naše namene, da bi pridobili takšne predstavitve zvočnih posnetkov, s katerimi bi izvajali razvrščanje posnetkov na govor in ne–govor samo na podlagi dveh modelov. Tudi primerjava poteka CVS značilk, izpeljanih iz dveh različnih razpoznavalnikov, zagotavlja učinkovitost ocen značilk v primeru različnih razpoznavalnikov glasov. Potek značilk se namreč v obeh primerih zelo dobro ujema, kar priča, da tudi različne govorne enote, ki jih uporabljamo v primeru slovenskega in angleškega razpoznavalnika glasov, ne vplivajo bistveno na izračun značilk. To pomeni, da so izbrane značilke tudi neodvisne od jezika.

V zaključku lahko povzamemo, da smo s predlaganimi fonetičnimi značilkami poskušali pridobiti takšne predstavitve signalov, s katerimi bi bilo mogoče ločevati govor od vseh ostalih ne–govornih pojavov. To smo naredili tako, da smo ocenjevali delovanje razpoznavalnikov glasov na podlagi samodejno pridobljenih transkripcij osnovnih govornih enot. Tako smo izpeljali štiri fonetične značilke, s katerimi smo merili, kako dobro delujejo razpoznavalniki v primeru govornih in ne–govornih posnetkov. Značilke so bile zasnovane tako, da so bile neodvisne od jezika razpoznavanja in modelov razpoznavalnika. Na ta način smo prenesli odločanje o govornih in ne–govornih pojavih v zvočnih posnetkih iz akustičnega nivoja na višji - fonetični nivo.

V nadaljevanju si bomo pogledali, kako smo predlagane značilke vključili v postopke GNG segmentacije, kako smo jih združevali z akustičnimi značilkami in kakšne rezultate segmentacije smo dosegli na različnih zbirkah zvočnih posnetkov.

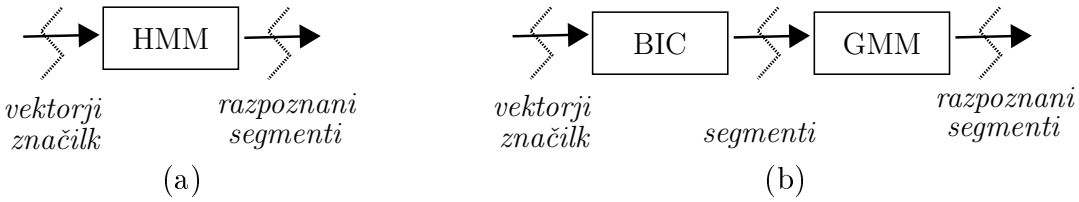
[2]Prikaz značilk je bil izveden z orodjem wavesurfer (http://www.speech.kth.se/wavesurfer/).

3.3 Segmentacija zvočnih posnetkov na govorne in ne–govorne dele

3.3.1 Postopki GNG segmentacije

Pri razvoju fonetičnih značilk smo preizkušali dva postopka GNG segmentacije [Žibert-06], ki sta shematično prikazana na sliki 3.3.

V prvem postopku (slika 3.3 (a)) se izvaja segmentacija in razvrščanje segmentov na govor in ne–govor hkrati, v drugem postopku (slika 3.3 (b)) pa sta bili fazi segmentacije in razvrščanja segmentov ločeni. V prvem primeru se je izvajala GNG segmentacija s pomočjo prikritih Markovovih modelov (HMM), v drugem primeru pa se je naprej izvedla segmentacija zvočnih posnetkov glede na akustične spremembe v signalih, nato so se na pridobljenih segmentih izračunale fonetične značilke, na podlagi katerih se je potem izvedlo razvrščanje na govor in ne–govor.



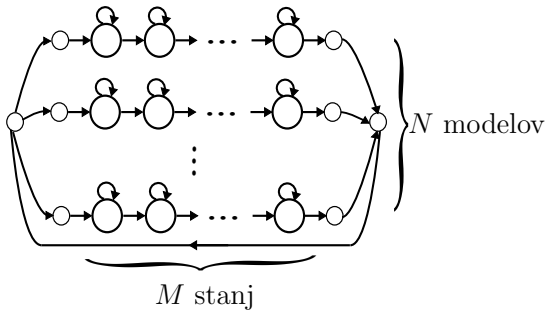
Slika 3.3: Shemi dveh postopkov GNG segmentacije. Pri shemi (a) se segmentacija in razvrščanje segmentov izvajata sprotno z uporabo HMM modelov in s postopkom Viterbijevega dekodiranja. Shema (b) prikazuje zaporeden postopek segmentacije in razvrščanja: v prvem koraku se izvede segmentacija na podlagi akustičnih predstavitev zvočnih posnetkov, v drugi pa razvrščanje segmentov s pomočjo GMM modelov.

V obeh primerih smo za razvrščanje posnetkov oziroma segmentov uporabljali GMM modele, ki smo jih pridobili iz učnega materiala govornih in ne–govornih posnetkov. Učenje GMM modelov je potekalo na standarden način z uporabo EM algoritma[3] [Theodoridis-03, str. 491–494].

V primeru prvega postopka iz slike 3.3 (a) smo sledili izvedbi GNG segmentacije s HMM modeli, ki je bila predstavljena v [Ajmera-04]. Razlika je bila le v tem, da so v [Ajmera-04] izvajali segmentacijo zvočnih posnetkov glede na govor in glasbo, mi pa smo postopek posplošili na detekcijo govornih in ne–govornih odsekov. Zato smo ta postopek obravnavali kot referenčni postopek, s katerim smo primerjali rezultate GNG segmentacije ob uporabi različnih predstavitev zvočnih posnetkov. Osnova za izgradnjo HMM modelov so bili GMM modeli, ki smo jih pridobili z učenjem na govornih in ne–govornih predstavitvah zvočnih posnetkov. Iz GMM modelov govora in ne–govora smo sestavili HMM modele, ki smo jih povezali v mrežo s povratno zanko.

Topologija HMM modelov je prikazana na sliki 3.4. Vsako stanje HMM mreže predstavlja en GMM model. Mreža je sestavljena iz N HMM modelov, ki predstavljajo

[3] EM algoritem je kratica za *Expectation-Maximization (EM) algorithm*, ki je bil prvič predstavljen v [Dampster-77].



Slika 3.4: Topologija HMM modelov, ki smo jih uporabljali pri GNG segmentaciji.

posamezne razrede govora in ne–govora. V primeru akustičnih predstavitev je bilo razredov več, npr. model čistega govora, telefonskega govora, model glasbe, šuma, tišine ipd. V primeru fonetičnih značilk pa smo uporabljali samo dva HMM modela: model govora in model ne–govora. HMM modeli so bili sestavljeni iz M stanj. Vsako stanje HMM modela je vsebovalo isti GMM model posameznega razreda. Z razmnoževanjem GMM modelov v M stanj predpišemo najmanjše možno trajanje odsekov, ki jih lahko modeliramo s takimi HMM modeli. Vrednosti povezav med stanji v HMM modelih so bile določene ročno, vrednosti povezav med HMM modeli, s katerimi določamo verjetnosti posameznih modelov, pa smo določali na podlagi optimalnih rezultatov GNG segmentacije na razvojni zbirki. Postopek določanja optimalnih vrednosti povezav bomo predstavili v naslednjem razdelku.

GNG segmentacija na podlagi takšnih HMM modelov je potekala podobno kot pri razpoznavanju govora s HMM modeli. S postopkom Viterbijevega dekodiranja [Rabiner-89] smo poskušali poiskati najbolj verjetno zaporedje govornih in ne–govornih HMM modelov, s katerimi bi najbolje opisali dano predstavitev signala, ki smo ga obdelovali. Rezultat je bilo zaporedje oznak HMM modelov, ki jim je bilo pridruženo trajanje posamezne oznake. Na ta način smo dobili segmentirane zvočne posnetke na govorne in ne–govorne dele.

Drugi postopek, prikazan na sliki 3.3 (b), smo razvili prav za namene GNG segmentacije s fonetičnimi značilkami. V tem postopku se je izvajala segmentacija in razvrščanje segmentov ločeno. V prvi fazi se je izvedla segmentacija glede na akustične lastnosti zvočnih posnetkov. Postopek segmentacije, ki smo ga uporabljali v našem primeru, bo podrobneje opisan v naslednjem poglavju. Bistvo segmentacije je, da na podlagi akustičnih značilk - v našem primeru smo uporabljali MFCC značilke - z uporabo kriterija BIC [Chen-98, Tritschler-99] najprej razdelimo zvočni posnetek na večje odseke glede na zamenjave govorcev in spremembe akustičnih ozadij. Na teh odsekih smo nato na podlagi glasovnih transkripcij izračunali fonetične značilke CVS (VUS). Tako smo za vsak odsek pridobili samo en vektor značilk (4 značilke). Na podlagi vsakega takega vektorja smo nato z GMM modelom določili, kateremu razredu je pripadal dani segment. GMM modeli, ki smo jih tu uporabljali, so bili isti kot v primeru GNG segmentacije s HMM modeli. Predlagana metoda GNG segmentacije je primerna za CVS (VUS) značilke, saj v tem primeru dobimo večje odseke signalov, na katerih lahko bolj zanesljivo ocenimo CVS (VUS) značilke kot pa v primeru krajših, vnaprej določenih odsekov. Predlagani postopek segmentacije smo primerjali s prvim postopkom samo v

primeru fonetičnih značilk.

3.4 Preizkusi postopkov segmentacije

V preizkusih GNG segmentacije smo preverjali dvoje: postopke segmentacije in predstavitve signalov, ki bi bile primerne za GNG segmentacijo zvočnih posnetkov informativnih oddaj. Pri tem nas je zanimala predvsem tista kombinacija postopkov in predstavitve zvočnih posnetkov, s katero bi dosegli najboljše rezultate GNG segmentacije v različnih pogojih delovanja, v različnih akustičnih situacijah in pri različnih tipih ne–govornih signalov. Dodatna zahteva pri GNG segmentaciji je bila, da bi poiskali takšen postopek GNG segmentacije, ki bi ga lahko enostavno vključili v sisteme nadaljnje obdelave zvočnih posnetkov. Zato smo dodatno ocenjevali tudi časovno in računsko zahtevnost primerjanih postopkov segmentacije.

Tako smo preizkušali tri tipe predstavitev zvočnih posnetkov in dva postopka segmentacije, ki smo ju že predstavili v prejšnjem razdelku. Pri predstavitvah signalov smo se omejili na naslednje skupine značilk:

- *akustične značilke*, ki smo jih opisovali s koeficienti melodičnega kepstra (MFCC);
- *entropijo in dinamizem*, ki sta ju predlagala Williams in Ellis [Williams-99] in sta bili že uspešno uporabljeni pri detekciji govora in glasbe [Ajmera-03];
- predlagane *CVS (VUS) značilke*, ki smo jih opisali v razdelku 3.2.2.

Dodatno smo izvedli tudi postopke segmentacije s kombinacijo vseh treh tipov značilk. Tako smo izvedli dva sistema fuzije GNG segmentacije, ki sta temeljila na kombinaciji MFCC in CVS značilk ter na kombinaciji entropije in dinamizma z MFCC značilkami. V nadaljevanju bomo tako podrobneje opisali pridobivanje vseh treh skupin značilk in natančneje opisali izvedbo postopkov GNG segmentacije, ki smo ju predstavili v prejšnjem razdelku.

3.4.1 Preizkušane predstavitve zvočnih posnetkov GNG segmentacije

Osnovni referenčni sistem GNG segmentacije je predstavljal postopek GNG segmentacije z uporabo HMM modelov, kjer so bili zvočni posnetki predstavljeni z MFCC značilkami. MFCC značilke skupaj z oceno kratkočasovne energije signala in z izvedbo prvih in drugih odvodov osnovnih značilk, ki jih izpeljemo iz koeficientov regresijskih premic, predstavljajo osnovno akustično parametrizacijo signalov v sistemih za razpoznavanje govora [Picone-93]. Uporabljajo pa se tudi v drugih postopkih govornih tehnologij. Izvedba značilk je narejena tako, da z njimi lahko dobro modeliramo osnovne enote govora, vendar se je izkazalo, da delujejo dobro tudi v primeru detekcije in ločevanja govora od ostalih ne–govornih pojavov [Carey-99]. Ravno zato smo se odločili, da bomo v osnovnem sistemu uporabljali 12 MFCC značilk, ki smo jim pridružili

še normalizirano[4] kratkočasovno energijo signala in prve odvode osnovnih značilk. Pri tem moramo omeniti, da smo v naših preizkusih uporabljali tudi druge odvode, vendar z njimi nismo izboljšali rezultatov GNG segmentacije.

Druga skupina značilk so bile značilke, ki so temeljile na meri entropije in dinamizma. V naših preizkusih smo uporabljali kratkočasovna povprečja obeh mer, ki so definirana v [Ajmera-03]. Z obema značilkama merimo delovanje preprostih razpoznavalnikov govora. V našem primeru smo uporabljali razpoznavalnik glasov, ki je temeljil na HMM modelih in na akustičnih predstavitvah signalov z MFCC značilkami. V tem primeru se tako izračun entropije in dinamizma prevede na ocenjevanje posteriornih verjetnosti posameznih stanj v določenem trenutku razpoznavalnika, ki so odvisne od zaporedja vektorjev MFCC značilk. Entropija nam predstavlja mero določenosti takšnega sistema; večja kot je, bolj razpršene so verjetnosti stanj, manjša kot je, bolj je sistem določen. To z drugimi besedami pomeni, večja kot je entropija, manjša je verjetnost, da je dan signal govor in obratno. Pri dinamizmu pa ocenjujemo razlike med verjetnostmi stanj pri prehodu med dvema zaporednima vektorjema MFCC značilk. Tu je situacija ravno obratna, večja kot je verjetnost, da ostajamo znotraj istega stanja v HMM modelu, bolj verjetno je dan signal govor. Razpoznavalnik glasov, ki smo ga uporabljali v naših eksperimentih za izračun entropije in dinamizma, smo zgradili na podlagi govornih podatkov iz zbirke TIMIT [Garofolo-93]. Razpoznavanje glasov je potekalo na standarden način z uporabo 12 MFCC značilk z energijo in prvimi in drugimi odvodi [Young-04]. Vse ostale parametre za izračun entropije in dinamizma pa smo povzeli po [Ajmera-04].

Osnova za izračun CVS (VUS) značilk so samodejne transkripcije, ki jih pridobimo iz razpoznavalnikov glasov. V naših preizkusih smo uporabljali dva razpoznavalnika: prvi je bil namenjen za razpoznavanje slovenskih glasov govora, drugi pa razpoznavanju angleškega govora. Slovenski razpoznavalnik glasov smo zgradili iz govornih podatkov treh slovenskih govornih zbirk: GOPOLIS, VNTV in K211d, [Mihelič-03]. Ta razpoznavalnik smo zato označili kot *razpoznavalnik–SI*. Drugi razpoznavalnik, ki smo ga označili kot *razpoznavalnik–EN*, pa je bil naučen na podlagi zbirke TIMIT [Garofolo-93]. Razpoznavalnik–EN je bil enak, kot razpoznavalnik glasov, ki smo ga uporabljali za izračun entropije in dinamizma. Oba sistema za razpoznavanje glasov sta bila zgrajena iz osnovnih govornih enot, ki smo jih modelirali s HMM modeli. Vsak HMM model je bil sestavljen iz treh stanj GMM modelov z diagonalnimi kovariančnimi matrikami. Ocenjevanje parametrov HMM modelov je potekalo s postopkom Baum–Welch na standarden način [Young-04]. Zaradi različnega jezika modeliranja smo v obeh primerih razpoznavalnikov modelirali različne skupine osnovnih govornih enot. V primeru razpoznavalnika–SI smo uporabljali 38 monofonskih enot govora, v primeru razpoznavalnika–EN pa 48 monofonov, ki smo jih pridobili iz osnovnih 68 enot po postopku opisanem v [Lee-89]. Topologija mreže HMM modelov obeh razpoznavalnikov je bila postavljena glede na bigramske jezikovne modele glasov, ki smo jih ocenili iz danih zbirk učenja. V fazi razpoznavanja smo v obeh primerih uporabljali običajno izvedbo MFCC značilk z energijo in prvimi in drugimi odvodi. Pri tem smo izvedli tudi analizo natančnosti razpoznavanja obeh razpoznavalnikov. Z razpoznavalnikom–

[4]Normalizacija energije je bila potrebna za boljšo detekcijo ne–govornih delov, predvsem tišine, v različnih akustičnih razmerah.

SI smo na testnem delu zbirke GOPOLIS dosegli 70% natančnost razpoznavanja glasov, z razpoznavalnikom–EN pa na testnem delu zbirke TIMIT 61% natančnost. Ker pa so nas v primeru CVS (VUS) značilk zanimale predvsem enote CVS (VUS), smo v primeru obeh razpoznavalnikov ocenili še natančnost razpoznavanja CVS (VUS) enot. V primeru razpoznavalnika–SI je bila natančnost 88%, v primeru razpoznavalnika–EN pa 75%. Podobne rezultate smo dobili tudi v primeru VUS enot.

Po prevedbi samodejnih transkripcij iz osnovnih govornih enot v CVS (VUS) enote smo izvedli izračun značilk CVS (VUS) po formulah iz (3.1) - (3.4). Pri izračunu značilke iz (3.1) je bila α v vseh primerih postavljena na 0.5. Prevedba transkripcij je potekala za vsak jezik razpoznavanja posebej. Pri tem smo ocenjevali vrednosti značilk na dva načina: na vnaprej določenih oknih analize in na akustičnih segmentih. Pri prvotnem preizkušanju fonetičnih značilk se je namreč izkazalo, da pridobimo bolj zanesljive ocene iz daljših odsekov zvočnih posnetkov. Ravno zaradi tega smo tudi razvili drugi postopek segmentacije, kjer s postopkom BIC najprej določimo akustične segmente [Chen-98, Tritschler-99] in na njih nato ocenimo fonetične značilke, na podlagi katerih z GMM modeli razvrščamo segmente na govor ali ne–govor. Zato smo ta postopek označili kot *BIC–GMM segmentacija*.

3.4.2 Določanje parametrov postopkov GNG segmentacije

Vse prej opisane predstavitve signalov zvočnih posnetkov smo kombinirali z dvema postopkoma GNG segmentacije, ki sta bila predstavljena v razdelku 3.3.

V primeru GNG segmentacije s HMM modeli (slika 3.3 (a)) je bilo potrebno vektorje značilk vseh treh predstavitev izračunavati na konstantno določenih oknih analize z vnaprej določenim premikom izračuna. V primeru MFCC značilk ter entropije in dinamizma je bil premik postavljen na 10 ms, v primeru CVS (VUS) značilk pa smo ocenjevali značilke na oknih dolžine 3.0 s na vsakih 100 ms. Na ta način smo v vseh treh primerih pridobili zaporedja vektorjev značilk, ki smo jih razvrščali v zaporedja govornih in ne–govornih oznak glede na HMM modele z dinamičnim postopkom Viterbijevega dekodiranja [Rabiner-89].

V primeru GNG segmentacije s postopkom *BIC–GMM segmentacije* je bilo potrebno najprej določiti akustične segmente. To smo storili s postopkom segmentacije BIC [Chen-98, Tritschler-99]. Segmentacija je potekala na podlagi 12 MFCC značilk z energijo, kjer smo določali meje med segmenti s kriterijem BIC na podlagi Gaussovih porazdelitev, ocenjenih z enim povprečnim vektorjem in polno kovariančno matriko. Prag za določitev meje smo postavili na razvojni zbirki, ki bo opisana v nadaljevanju. Podrobnejši opis postopka akustične segmentacije s kriterijem BIC bo predstavljen v naslednjem poglavju. Razvrščanje segmentov na govor in ne–govor je potekalo z uporabo GMM modelov. Segment je bil razvrščen v enega izmed razredov glede na največjo verjetnost GMM modela, s katerim smo opisali dani segment. Postopek *BIC–GMM segmentacije* smo uporabljali samo v primeru fonetičnih značilk.

Osnova za določitev govornih in ne–govornih odsekov v obeh postopkih GNG segmentacije so bili GMM modeli. V vseh primerih segmentacije smo Gaussove porazdelitve

v GMM modelih opisovali z diagonalnimi kovariančnimi matrikami, parametre pa smo določali s postopkom EM [Theodoridis-03, str. 491–494]. V primeru MFCC značilk ter značilk entropije in dinamizma smo uporabljali več modelov za modeliranje govora in več za modeliranje ne–govora. Govor smo modelirali z dvema modeloma, in sicer glede na kanal posnetka (telefon, ne–telefon). Razred ne–govora pa je bil prav tako sestavljen iz dveh GMM modelov: prvi je predstavljal glasbo, drugi pa premore. V primeru CVS (VUS) predstavitev smo vsak razred modelirali samo z enim modelom. Tako je en GMM model predstavljal govor, drugi pa ne–govor. Vsi modeli so bili naučeni na podlagi podatkov iz učnih zbirk. Število Gaussovih porazdelitev v vsakem GMM modelu pa je bilo določeno na podlagi optimalnih rezultatov GNG segmentacije na razvojni zbirki. V primeru MFCC značilk smo uporabljali GMM modele sestavljene iz 128 Gaussovih porazdelitev. V primeru značilk entropije in dinamizma smo določili 4 porazdelitve na model (v [Ajmera-03] so uporabljali samo 2). V primeru CVS (VUS) značilk pa je bil vsak model sestavljen samo iz dveh Gaussovih porazdelitev. Tako določene in naučene GMM modele smo uporabili v obeh primerih postopkov GNG segmentacije.

V primeru segmentacije s HMM modeli smo morali dodatno določati stanja modelov in nastavljati vrednosti povezav v HMM mreži, ki predstavljajo prehodne verjetnosti med posameznimi modeli. Tu smo se zgledovali po nastavitvah, ki so bile opisane v [Ajmera-03]. Za določitev vseh odprtih parametrov smo uporabljali razvojno zbirko. Ker smo ugotovili, da v zbirki ni govornih in ne–govornih segmentov krajših od 1.4 s, smo ustrezno načrtovali tudi HMM modele. Tako smo v primeru značilk MFCC ter entropije in dinamizma določili 140 stanj HMM modelov, kar je ob izračunu teh značilk na vsakih 10 ms ustrezalo ravno trajanju 1.4 s. Podobno smo tudi v primeru CVS (VUS) značilk določili 14 stanj zaradi izračuna značilk na vsakih 100 ms. Vse prehodne verjetnosti povezav v HMM modelih smo nastavili na 0.5. Vrednosti povezav med HMM modeli pa so bile določene tako, da smo favorizirali enega izmed razredov glede na optimalne rezultate GNG segmentacije na razvojni zbirki. Postopek in izbira teh uteži bodo predstavljeni v naslednjih razdelkih.

Postopke GNG segmentacije z uporabo HMM modelov smo izvajali z orodji iz zbirke HTK Toolkit [Young-04]. Za določitev parametrov GMM modelov (EM postopek, učenje) in izvajanje segmentacije s postopkom BIC (pri *BIC–GMM segmentaciji*) pa smo razvili svoja lastna orodja.

3.4.2.1 Računska zahtevnost postopkov GNG segmentacije

Jasno je, da je izvedba značilk, ki temeljijo na opisovanju delovanja razpoznavalnikov govora, računsko in časovno zahtevnejša, kot pa sam izračun akustičnih značilk. Zato nas je v primeru fonetičnih značilk in značilk entropije in dinamizma zanimalo, kako se poveča računska zahtevnost postopkov GNG segmentacije.

V našem primeru smo za izračun obeh skupin značilk (entropije–dinamizma in CVS (VUS) značilk) uporabljali sorazmeroma preproste razpoznavalnike glasov, s katerim smo pospešili proces razpoznavanja govora. Kljub temu je pridobivanje takih značilk ravno zaradi procesa razpoznavanja mnogo bolj zahtevno, kot pa pridobivanje samo

akustičnih predstavitev. Iz poskusov smo ocenili, da je časovna zahtevnost procesa izračuna CVS (VUS) značilk približno 3-krat večja kot v primeru MFCC značilk. Tako smo npr. za izračun CVS značilk v primeru razpoznavalnika–SI s standardnim PC računalnikom potrebovali v povprečju približno 25% časa dolžine zvočnega posnetka, ki smo ga obdelovali. Po drugi strani je postopek GNG segmentacije s HMM modeli pri CVS (VUS) značilkah potekal hitrejše, saj smo v tem primeru potrebovali samo dva modela za razvrščanje segmentov. V primeru ostalih značilk smo imeli več modelov. Vendar je bila razlika v času delovanja zanemarljiva, saj smo v vseh primerih izvajali postopke Viterbijevega dekodiranja na sorazmerno majhnem številu HMM modelov. Na hitrost GNG segmentacije pa je predvsem vplival korak izračuna vektorjev značilk. V primeru značilk MFCC in entropije–dinamizma je bil korak 10-krat manjši kot v primeru CVS (VUS) značilk, kar je pomenilo hitrejšo GNG segmentacijo v zadnjem primeru.

Dodaten problem je predstavljal tudi način izračuna fonetičnih značilk. Zvočni posnetek je bilo potrebno namreč najprej obdelati z razpoznavalnikom in šele nato izračunati značilke. To pomeni, da takšen sistem deluje v dveh fazah. V prvi fazi poteka razpoznavanje, v drugi pa odločanje o govornih in ne–govornih odsekih. Podobno lahko ugotovimo tudi za postopek *BIC–GMM segmentacije*: najprej je potrebno izvesti segmentacijo in šele nato razvrščanje. Vendar z nekaj izboljšavami v obeh primerih lahko prevedemo postopke, da delujejo samo v enem koraku, kar je priporočljivo za vključevanje takšnih postopkov v sisteme nadaljnje obdelave zvočnih posnetkov, ki delujejo v stvarnem času. Tako lahko npr. daljši zvočni posnetek razdelimo na krajše odseke, ki jih nato obdelamo s predlaganimi postopki GNG segmentacije. V tem primeru sicer dodamo določeno zakasnitev v sistem, ki pa jo lahko prilagajamo glede na tip in namen aplikacije, ki jo izvajamo.

3.4.3 Združevanje predstavitev zvočnih posnetkov pri GNG segmentaciji

Kot smo že omenili, smo želeli pridobiti čimbolj robustne in zanesljive predstavitve zvočnih posnetkov, ki bi bile primerne za GNG segmentacijo v različnih pogojih delovanja, zato smo v okviru našega raziskovalnega dela preizkušali tudi kombinacije vseh treh predlaganih predstavitev. Osnovna ideja je bila, da bi akustična informacija v signalu predstavljala osnovno informacijo za GNG segmentacijo, z informacijo višjega reda izpeljano iz akustičnih predstavitev (entropija–dinamizem, fonetične značilke) pa bi zagotavljali večjo neobčutljivost sistemov GNG segmentacije.

Zaradi tega smo izvedli dve skupini združevanja predstavitev: v prvi skupini smo združevali MFCC značilke z entropijo in dinamizmom, v drugi skupini pa MFCC in fonetične značilke. V obeh primerih je bilo potrebno časovno uskladiti različne predstavitve zvočnih posnetkov. To smo dosegli tako, da smo v vseh treh skupinah predstavitev pridobivali vektorje značilk ob enakih časovnih premikih izračunov. Tako smo v obeh primerih združevanja dobili dva toka predstavitev *(ang. stream)*, ki smo ju združevali v modelih GMM. GNG segmentacija z združevanjem predstavitev je potekala samo v primeru uporabe HMM modelov, ki smo jih tvorili na podoben način, kot je bilo že

opisano. Razlika je bila le v tem, da smo v stanjih HMM modelov vodili GMM modele, sestavljene iz dveh tokov predstavitev. Odločitev razvrščanja na govor in ne–govor je tako potekala na podlagi kombinacije verjetnosti iz obeh tokov predstavitev v vsakem stanju HMM modela. V našem primeru smo se odločili za združevanje odločitev na podlagi utežene vsote logaritma verjetnosti posameznih tokov predstavitev. Tak postopek združevanja različnih tokov predstavitev v HMM modelih je bil predlagan v [Potamianos-04]. V primeru GNG segmentacije smo te postopke označili kot *GNG segmentacija s fuzijo*.

V primeru takšne GNG segmentacije smo združene GMM modele sestavili iz osnovnih GMM modelov posamičnih predstavitev, ki smo jih pridobili s postopkom EM. Dodatno pa smo morali pri tem določati še uteži združevanja, ki smo jih pridobili na podlagi optimalnih rezultatov GNG segmentacije na razvojni zbirki.

3.4.4 Podatkovne zbirke zvočnih posnetkov za vrednotenje postopkov GNG segmentacije

Podatkovne zbirke, ki smo jih uporabljali za vrednotenje postopkov GNG segmentacije, so bile sestavljene iz zvočnih posnetkov informativnih oddaj iz zbirk SiBN in COST278. Razvojno zbirko pa smo sestavili iz posebno izbranih posnetkov različnih TV oddaj. Glavni namen eksperimentalnih zbirk je bil, da bi zagotovili čimbolj pestro akustično in jezikovno vsebino zvočnih posnetkov, s katerimi bi zajemali različne govorne in ne–govorne pojave. Tako bi lahko ocenjevali postopke GNG segmentacije in predstavitve zvočnih posnetkov v različnih pogojih delovanja in v različnih govornih in ne–govornih situacijah.

Razvojna zbirka je vsebovala 3 ure zvočnih posnetkov dveh zabavnih TV oddaj. Prva oddaja (2 uri) je bila v slovenskem jeziku, druga (1 ura) pa v italijanskem. Posnetki so bili izbrani tako, da so vsebovali približno 2/3 govora, preostala tretjina pa je pripadala različnim ne–govornim situacijam. Ker smo izbrane posnetke pridobili iz zabavnih oddaj, je večina ne–govornih pojavov pripadala glasbi, različnim aplavzom, smehu, veliko je bilo tudi tišine, raznih zvočnih efektov ipd. Govor v posnetkih je prispevalo več različnih govorcev, ki so govorili v slovenskem in italijanskem jeziku. Zaradi narave oddaj je bil govor v posnetkih večinoma spontan z veliko nejezikovnimi elementi, ki smo jih v glavnem opredelili kot govor.
Razvojno zbirko smo uporabljali za nastavitev vseh odprtih parametrov postopkov in predstavitev, predvsem pa za določitev optimalnih uteži modelov detekcije GNG razpoznavanja, ki smo jih kasneje uporabljali pri testnih zbirkah.

Ostali eksperimentalni zbirki sta bili sestavljeni iz posnetkov informativnih oddaj iz zbirk SiBN in COST278, ki smo ju že opisali v prejšnjem poglavju. S stališča govornih in ne–govornih elementov, ki jih vsebujejo informativne oddaje, lahko ugotovimo, da je bila skupna značilnost vseh posnetkov, da je v njih prevladoval govor, ne–govorni elementi pa so pripadali predvsem glasbi najavnih in/ali odjavnih TV špic, premorom med posameznimi novicami in različnim šumom v ozadju TV poročil. Govor je bil sestavljen iz velikega števila različnih govorcev, ki govorijo v različnih jezikih, v različnih akustičnih situacijah in na različne načine. Kot smo že omenili v prejšnjem poglavju, je

bila zbirka COST278 zaradi načina pridobivanja posnetkov informativnih oddaj veliko bolj pestra z različnimi govornimi in ne–govornimi pojavi v primerjavi z zbirko SiBN.

Podatke vseh treh eksperimentalnih zbirk smo razdelili na učni in testni del. Učne posnetke so predstavljale 3 oddaje iz vsake zbirke v skupnem trajanju okoli treh ur. Te posnetke smo uporabili za učenje GMM modelov GNG segmentacije v vseh primerih predstavitev. Testni del razvojne zbirke (2 uri) smo uporabljali za določitev vseh ostalih parametrov postopkov: uteži modelov detekcije, uteži GMM modelov pri fuziji, nastavitev pragov za BIC segmentacijo in nastavitev optimalnih parametrov za pridobivanje CVS (VUS) značilk. Testni del zbirke SiBN je vseboval 30 ur posnetkov informativnih oddaj, testni del zbirke COST278 pa 25 ur. Oba sta služila za primerjavo in vrednotenje postopkov GNG segmentacije.

Natančnejša razdelitev posnetkov na učne, razvojne in testne množice, ki smo jih uporabljali pri preizkusih postopkov GNG segmentacije, je opisana v dodatku A disertacije.

3.4.5 Mere vrednotenja postopkov GNG segmentacije

Pri vrednotenju postopkov GNG segmentacije smo merili natančnost razpoznavanja glede na čas skupnega ujemanja detektiranih govornih in ne–govornih segmentov z referenčnimi segmenti. Pri tem smo uporabljali tri mere: delež ujemanja govornih segmentov, delež ujemanja ne–govornih segmentov in skupni delež ujemanja obeh tipov segmentov. S prvo mero smo tako merili pravilno *razpoznavanje govora*, z drugo pravilno *razpoznavanje ne–govora*, s tretjo pa *skupno natančnost razpoznavanja (skupno razpoznavanje)*. Čas ujemanja smo merili glede na način izračuna vektorjev značilk posameznih predstavitev. Tako smo v primeru značilk MFCC in entropije–dinamizma čas ujemanja zaokroževali na 10 ms natančno, v primeru CVS (VUS) značilk pa na 100 ms.

Pri tem moramo omeniti, da je potrebno pri vrednotenju postopkov GNG segmentacije upoštevati vse tri mere in ne samo skupno natančnost razpoznavanja. To pa predvsem zato, ker se lahko zgodi, da je lahko en razred bolj izrazito zastopan v testnih podatkih in je skupna natančnost razpoznavanja bolj odvisna samo od razpoznavanja tega razreda. To se je zgodilo v primeru obeh testnih zbirk SiBN in COST278. V obeh primerih imamo namreč v povprečju okoli 90% govora in samo 10% ne–govora. To pa bi pomenilo, da bi lahko dosegli skupno natančnost razpoznavanja 90%, če bi celoten testni material razpoznavali kot govor. Na ta način seveda s skupno mero natančnosti ne bi mogli oceniti delovanje posameznih postopkov.
Pri razvojni zbirki smo kljub temu optimirali rezultate GNG segmentacije glede na skupno natančnost razpoznavanja, saj je bilo v tem primeru razmerje med govornimi in ne–govornimi podatki bolj uravnoteženo.

3.4.6 Primerjava postopkov GNG segmentacije na razvojni zbirki

Razvojno zbirko, sestavljeno iz treh ur posnetkov TV zabavnih oddaj, smo uporabili za dve vrsti preizkusov: za določanje optimalnih parametrov predstavitev in postopkov

GNG segmentacije ter za vrednotenje in izbiro optimalnih CVS (VUS) značilk, ki smo jih kasneje uporabljali za GNG segmentacijo zvočnih posnetkov iz zbirk SiBN in COST278.

V prvi skupini preizkusov smo določali takšne parametre predstavitev in modelov, da smo z njimi dosegli optimalne rezultate GNG segmentacije. Kriterij za določitev parametrov je bila *mera skupnega razpoznavanja*, ki smo jo predstavili v prejšnjem razdelku. Tu smo se predvsem ukvarjali z izbiro uteži modelov govora in ne–govora. Na slikah 3.5 in 3.6 so tako prikazani poteki rezultatov GNG segmentacije z različnimi postopki ob izbiri različnih kombinacij uteži modelov govora in ne–govora. Uteži modelov smo izbirali tako, da smo z njimi uravnavali skupno verjetnost razpoznavanja posameznih modelov govora in ne–govora, ki smo jih modelirali z GMM modeli. Na ta način smo v bistvu favorizirali en model (razred) proti drugemu. Z izbiro različnih uteži smo tako pridobili tudi različne rezultate GNG segmentacije. Optimalna kombinacija uteži je bila tista, pri kateri smo z določeno metodo dosegli najboljše rezultate skupnega razpoznavanja. Takšno kombinacijo smo poimenovali *optimalna izbira* parametrov in v eksperimentih s testnimi zbirkami smo tako primerjali postopke ob optimalnih in neoptimalnih izbirah parametrov.

Predstavitve in postopke GNG segmentacije, ki smo jih preizkušali in so prikazani tudi na slikah 3.5 in 3.6, lahko razdelimo na štiri skupine:

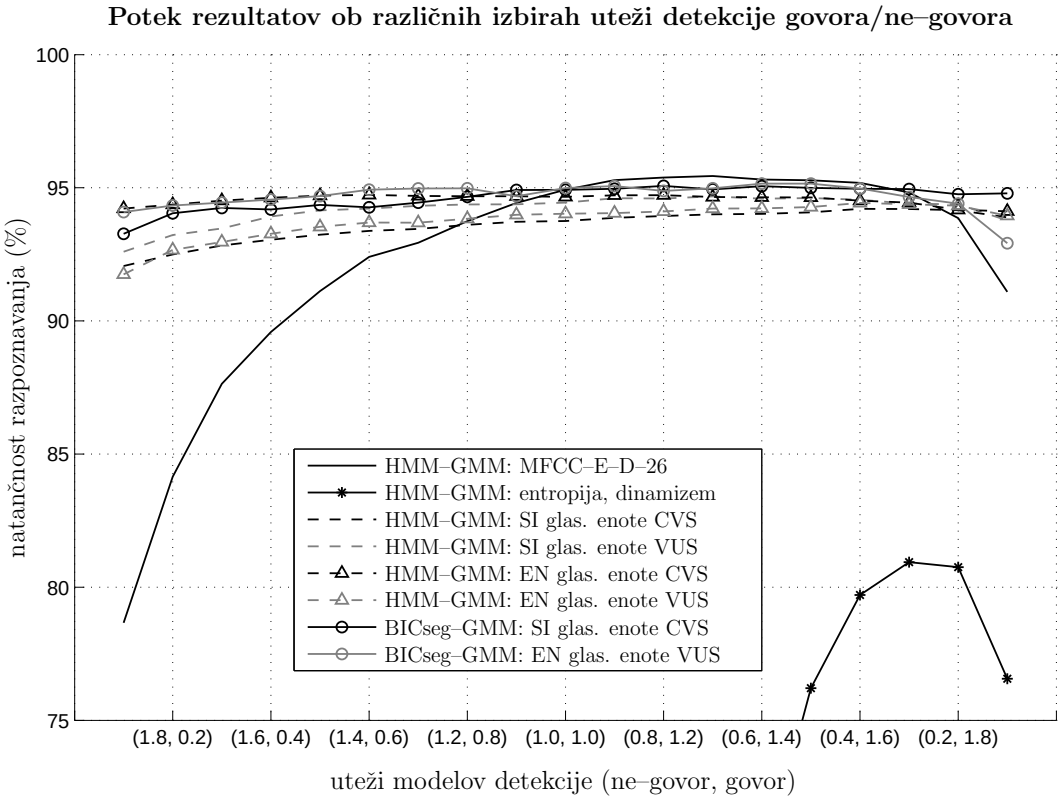
- MFCC značilke z GMM modeli:
  12 MFCC značilk z energijo in prvimi odvodi; GMM s 128 Gaussovimi porazdelitvami na model. Predstavitev smo označili kot *MFCC-E-D-26*.

- Značilke entropije in dinamizma z GMM modeli:
  povprečna entropija in dinamizem stanj HMM modelov razpoznavalnika–EN; GMM s 4-imi kombinacijami normalnih porazdelitev. Predstavitev smo označili kot *entropija, dinamizem*.

- Fonetične značilke CVS, VUS:
  značilke smo pridobili iz transkripcij CVS in VUS enot na podlagi razpoznavalnika–EN in razpoznavalnika–SI iz izrazov (3.1) - (3.4). Modelirali smo jih z GMM modeli z dvema normalnima porazdelitvama na model. Predstavitve smo označili kot *SI glas. enote CVS* in *SI glas. enote VUS* v primeru razpoznavalnika–SI in *EN glas. enote CVS* in *EN glas. enote VUS* v primeru razpoznavalnika–EN.

- Kombinacije predstavitev:
  v enem primeru je bila fuzija MFCC značilk in značilk entropija–dinamizem (označeno kot *fuzija MFCC+ent.,din.*), v drugem pa bila fuzija izvedena na podlagi značilk MFCC in CVS (označeno kot *fuzija MFCC+CVS*). CVS značilke smo pridobili na podlagi razpoznavalnika-SI na podoben način kot v prejšnjem primeru.

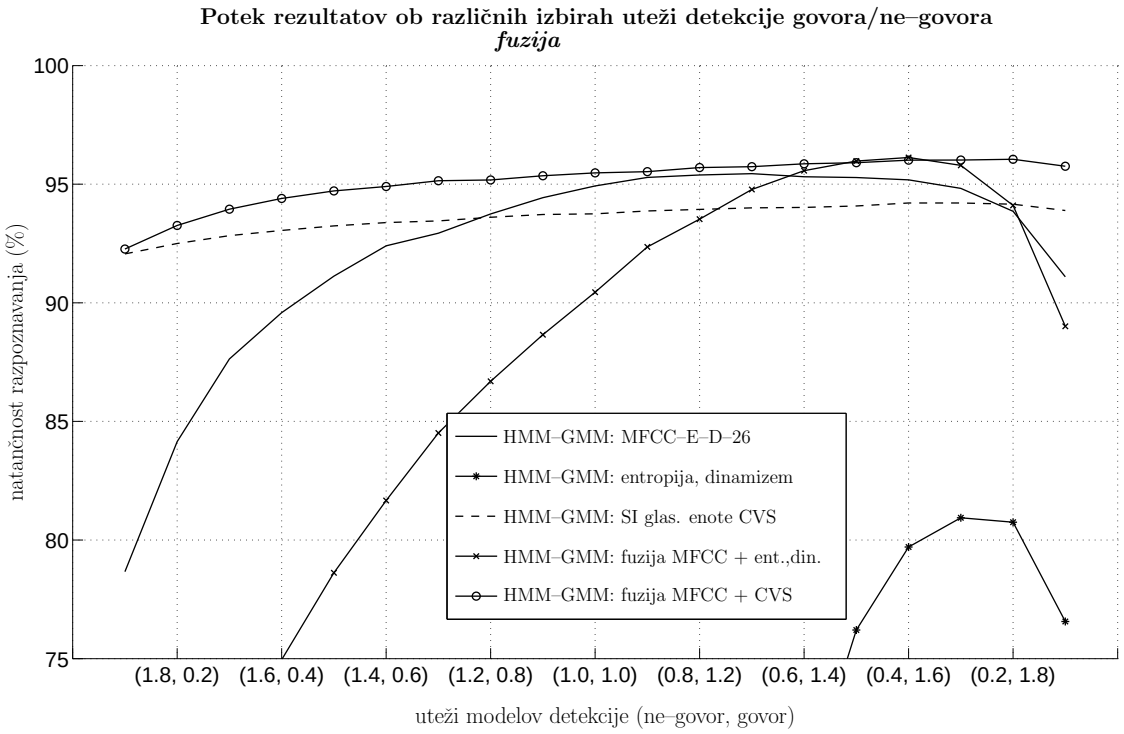
Preizkušali smo dva načina postopkov segmentacije. Vse predstavitve smo izpeljali s postopkom GNG segmentacije z uporabo HMM modelov, ki so bili sestavljeni iz GMM modelov. Te postopke smo označevali kot segmentacija *HMM–GMM*. Samo s



Slika 3.5: Določanje uteži modelov detekcije (ne–govor, govor) različnih postopkov glede na optimalne rezultate razpoznavanja na razvojni zbirki.

fonetičnimi predstavitvami pa smo testirali tudi drugi postopek, kjer je bila naprej izvedena BIC segmentacija in nato postopek razvrščanja s pomočjo GMM modelov. Ta postopek smo označevali kot segmentacija *BICseg–GMM*.

Na sliki 3.5 so prikazani poteki rezultatov skupnega razvrščanja različnih predstavitev GNG segmentacije. Primerjava rezultatov ob različnih izbirah uteži modelov detekcije pokaže bistveno razliko med fonetičnimi in ostalimi referenčnimi predstavitvami GNG segmentacije. Izkaže se namreč, da fonetične predstavitve delujejo stabilno na celotnem področju izbire uteži, medtem ko z ostalima dvema predstavitvama (MFCC in entropija–dinamizem) dosežemo najboljše rezultate na ozkih področjih izbire uteži. To pomeni bistveno prednost fonetičnih predstavitev, saj ne glede na optimalne izbire odprtih parametrov postopkov dosežemo dobre rezultate GNG segmentacije. Z drugimi besedami to pomeni, da so fonetične značilke manj občutljive na različne spremembe v delovanju postopkov in s tem tudi na spremembe kvalitete zvočnih posnetkov. To smo še dodatno pokazali v primeru testnih zbirk, ko smo izvajali preizkuse z optimalnimi in neoptimalnimi izbirami parametrov in smo s fonetičnimi značilkami dosegli znatno boljše rezultate kot pa z značilkami MFCC ter entropijo–dinamizmom. Na sliki lahko tudi vidimo, da z obema postopkoma GNG segmentacije (HMM–GMM, BICseg–GMM) v primeru fonetičnih značilk dosežemo podobne rezultate, ki se gibljejo v intervalu med 92% in 95% skupne natančnosti razpoznavanja. Iz rezultatov s slike 3.5 pa ne moremo sklepati o razliki GNG segmentacije v primeru CVS in VUS značilk, ki smo jih pridobili iz dveh različnih razpoznavalnikov glasov. Najboljše rezultate GNG segmentacije (ne-



Slika 3.6: Določanje uteži modelov detekcije (ne–govor, govor) različnih postopkov fuzije glede na optimalne rezultate razpoznavanja na razvojni zbirki.

kaj nad 95%) pa dosežemo v primeru MFCC značilk z utežmi ne–govornih modelov 0.8 in utežmi govornih modelov 1.2. S predstavitvami zvočnih posnetkov, ki so temeljile na značilkah entropije in dinamizma, smo na razvojni zbirki dosegli najslabše rezultate. Skupna ugotovitev na podlagi rezultatov na sliki 3.5 je, da ne glede na izbiro fonetičnih značilk dosežemo bolj stabilno delovanje postopkov GNG segmentacije v primerjavi z značilkami entropije in dinamizma in MFCC značilkami, s katerimi pa smo na razvojni zbirki dosegli najboljši rezultat GNG razpoznavanja.

Na sliki 3.6 so zbrani rezultati postopkov GNG segmentacije s fuzijo predstavitev. Podana je primerjava fuzije z značilkami MFCC in CVS (*MFCC+CVS*) in značilkami MFCC in entropije–dinamizma (*MFCC+ent.,din.*). Na sliki so prikazani še rezultati posamičnih predstavitev, s katerimi smo tvorili kombinacije predstavitev fuzije. Glavni namen fuzije predstavitev je bil, da bi združili akustične značilke, ki jih dobimo neposredno iz signalov zvočnih posnetkov, z izpeljanimi značilkami, ki jih tvorimo na podlagi delovanja sistemov za razpoznavanje govora. Kot je razvidno na sliki 3.6, smo na ta način izboljšali rezultate GNG segmentacije v obeh kombinacijah značilk. Podobno kot v prejšnjem primeru pa lahko ugotovimo, da je delovanje postopkov GNG segmentacije v primeru fuzije *MFCC+CVS* enako stabilno oziroma neodvisno od izbire uteži modelov detekcije. Pri fuziji *MFCC+ent.,din.* pa se je ta odvisnost še povečala, kar je predvsem posledica dejstva, da sta obe posamični predstavitvi, iz katerih je sestavljena združena predstavitev, močno odvisni od izbire uteži. Sklepna ugotovitev na podlagi rezultatov s slike 3.6 je tako, da smo s fuzijo predstavitev izboljšali rezultate GNG segmentacije in v primeru kombinacije MFCC značilk s predlaganimi fonetičnimi

značilkami CVS še izboljšali stabilnost delovanja postopkov.

Drugi namen razvojne zbirke je bil, da bi jo uporabili za oceno učinkovitosti delovanja posameznih fonetičnih značilk za detekcijo govornih in ne–govornih odsekov. Tako so v tabeli 3.1 zbrani rezultati razpoznavanja govora in ne–govora v primeru posameznih CVS značilk, ki smo jih izpeljali iz samodejnih transkripcij razpoznavalnika-SI. GNG segmentacija je bila izvedena z referenčnim postopkom HMM–GMM, uteži modelov detekcije pa so bile enake.

Tabela 3.1: Primerjava rezultatov GNG razpoznavanja z različnimi CVS značilkami iz (3.1) - (3.4). Primerjava je izvedena na razvojni zbirki in podana skupaj z rezultati ob uporabi vseh CVS značilk skupaj in uporabi MFCC značilk.

tip značilke	*razpoznavanje govora*	*razpoznavanje ne–govora*	*skupno razpoznavanje*
norm. raz. trajanja CV enot	82.3	70.0	77.8
norm. CV hitrost govora	89.6	93.7	91.1
norm. spremembe CVS enot	91.6	92.5	92.0
norm. raz. povp. trajanja CV enot	81.7	70.0	77.4
vse značilke CVS	94.7	93.4	94.2
značilke MFCC	93.5	97.4	94.9

Glede na rezultate razpoznavanja iz tabele 3.1 lahko ugotovimo, da z vsako od CVS značilk lahko dovolj zanesljivo detektiramo govorne in ne–govorne segmente. Z značilkami, ki temeljijo na spremembah CVS enot (normirana CV hitrost govora, normirane spremembe CVS enot), smo dosegli boljše rezultate razpoznavanja v primerjavi z značilkami, ki temeljijo na trajanju CVS enot (normirano razmerje trajanja CV enot, normirana razlika povprečnega trajanja CV enot). Preizkušali smo tudi vse preostale kombinacije CVS značilk, vendar v nobenem primeru nismo presegli rezultatov GNG segmentacije ob uporabi vseh štirih značilk. Zato smo tudi pri vrednotenju postopkov GNG segmentacije na testnih zbirkah uporabljali fonetične predstavitve sestavljene iz vseh štirih mer CVS (VUS) značilk.

Vzporedno z rezultati GNG razpoznavanja s posameznimi značilkami smo ugotavljali tudi stopnjo korelacije med posameznimi značilkami. Večje ujemanje smo tako izmerili med značilkami trajanja (normirano razmerje trajanja CV enot in normirana razlika povprečnega trajanja CV enot) in značilkami sprememb CVS enot (normirana CV hitrost govora, normirane spremembe CVS enot), kar je bilo zaradi izvedbe značilk pričakovano.

3.4.7 Primerjava postopkov GNG segmentacije na testnih zbirkah

Obsežnejši preizkusi postopkov GNG segmentacije so bili izvedeni s testnima zbirkama SiBN in COST278. Rezultati različnih postopkov in predstavitev GNG segmentacije so

zbrani v tabeli 3.2 za zbirko SiBN in v tabeli 3.3 v primeru zbirke COST278. V preizkusih na testnih zbirkah smo izvajali dve skupini primerjav. V prvi skupini preizkusov smo primerjali postopke GNG segmentacije ob izbiri optimalnih parametrov in uteži modelov detekcije, ki smo jih določili na razvojni zbirki. V drugi skupini preizkusov pa smo primerjali postopke segmentacije ob uporabi enakih uteži modelov detekcije, torej v primeru neoptimalno določenih parametrov postopkov. Na ta način smo želeli primerjati delovanje postopkov tudi v primeru, ko nimamo na razpolago razvojnih zbirk in je zato potrebno izvajati postopke GNG segmentacije ob neoptimalnih pogojih delovanja. Tako so prvi rezultati v tabelah 3.2 in 3.3 pridobljeni v primeru optimalno določenih uteži, rezultati v okroglih oklepajih () pa v primeru enakih uteži modelov detekcije.

Tabela 3.2: Rezultati GNG segmentacije na zbirki SiBN. Vrednosti v okroglih oklepajih () predstavljajo rezultate ob izbiri neoptimalnih vrednosti uteži modelov (enake uteži). Poudarjeni so najboljši rezultati v primeru fuzije in brez fuzije.

način razpoznavanja & tip značilk	*razpoznavanje govora*	*razpoznavanje ne–govora*	*skupno razpoznavanje*
HMM–GMM: MFCC	97.9 (96.4)	58.7 (72.3)	95.3 (94.8)
HMM–GMM: entropija, dinamizem	99.3 (88.9)	55.8 (88.7)	96.5 (88.9)
HMM–GMM: SI glas. enote, CVS	**98.2 (97.6)**	**91.1 (93.0)**	**97.8 (97.3)**
HMM-GMM: SI glas. enote, VUS	98.1 (97.7)	88.73 (90.1)	97.5 (97.2)
HMM–GMM: EN glas. enote, CVS	98.5 (98.4)	88.2 (88.8)	97.8 (97.7)
HMM-GMM: EN glas. enote, VUS	97.52 (96.7)	89.95 (92.9)	97.0 (96.4)
BIC–GMM: SI glas. enote, CVS	97.9 (97.9)	89.5 (89.7)	97.4 (97.3)
BIC–GMM: EN glas. enote, CVS	98.3 (98.2)	89.2 (89.2)	97.7 (97.7)
BIC-GMM: EN glas. enote, VUS	98.05 (97.9)	89.72 (90.2)	97.5 (97.4)
HMM–GMM: fuzija MFCC+ent.,din.	99.7 (97.9)	62.9 (88.9)	97.3 (97.3)
HMM–GMM: fuzija MFCC+SI-CVS	**99.3 (98.3)**	**87.0 (93.6)**	**98.5 (98.0)**

Kljub temu da sta zbirki SiBN in COST278 konceptualno različni in se zvočni posnetki informativnih oddaj med zbirkama razlikujejo po akustični, vsebinski in jezikovni vsebini, se rezultati GNG segmentacije v tabelah 3.2 in 3.3 ujemajo. To je predvsem posledica dejstva, da smo v obeh primerih uporabljali modele in nastavitve parametrov, ki smo jih določili iz enakih učnih in razvojnih podatkov. Skupna ugotovitev primerjave rezultatov iz obeh zbirk je, da smo s CVS (VUS) značilkami dosegli boljše rezultate GNG segmentacije kot pa z značilkami MFCC in značilkami entropije in dinamizma. Učinkovitost predlaganih fonetičnih značilk je še bolj izrazita, če primerjamo rezultate na podlagi ločenega razpoznavanja govora in ne–govora. V vseh primerih CVS (VUS) značilk lahko opazimo izrazito odstopanje rezultatov razpoznavanja ne–govornih segmentov v primerjavi z značilkami MFCC in entropije–dinamizma. To je ob podobnih rezultatih razpoznavanja govornih segmentov pomenilo boljše rezultate skupnega razpoznavanja. Še večje razlike v delovanju se pokažejo, če primerjamo delovanje GNG postopkov v primeru enakih uteži modelov detekcije. Tu pride do podobnega pojava kot v primeru razvojne zbirke. Izkaže se namreč, da se rezultati GNG segmentacije

Tabela 3.3: Rezultati GNG segmentacije na zbirki COST278. Vrednosti v okroglih oklepajih () predstavljajo rezultate ob izbiri neoptimalnih vrednosti uteži modelov (enake uteži). Poudarjeni so najboljši rezultati v primeru fuzije in brez fuzije.

način razpoznavanja & tip značilk	*razpoznavanje govora*	*razpoznavanje ne–govora*	*skupno razpoznavanje*
HMM–GMM: MFCC	98.7 (97.8)	44.0 (54.2)	94.6 (94.6)
HMM–GMM: entropija, dinamizem	98.5 (83.4)	38.4 (79.3)	94.0 (83.1)
HMM–GMM: SI glas. enote, CVS	96.6 (95.6)	76.9 (79.3)	95.1 (94.3)
HMM-GMM: SI glas. enote, VUS	97.2 (96.6)	72.2 (74.3)	95.3 (95.0)
HMM–GMM: EN glas. enote, CVS	97.9 (97.8)	71.1 (71.6)	95.9 (95.8)
HMM-GMM: EN glas. enote, VUS	96.8 (96.6)	72.4 (74.3)	95.0 (95.0)
BIC–GMM: SI glas. enote, CVS	97.1 (97.0)	76.3 (76.4)	95.6 (95.5)
BIC–GMM: EN glas. enote, CVS	**98.1 (98.0)**	**75.0 (75.2)**	**96.4 (96.3)**
BIC-GMM: EN glas. enote, VUS	97.7 (97.5)	75.2 (75.6)	96.0 (95.9)
HMM–GMM: fuzija MFCC+ent.,din.	99.4 (97.1)	34.7 (65.6)	94.6 (94.8)
HMM–GMM: fuzija MFCC+SI-CVS	**98.6 (97.0)**	**70.5 (78.4)**	**96.5 (95.6)**

v primeru značilk MFCC in entropije–dinamizma močno spreminjajo glede na izbiro uteži delovanja in kar je še slabše, boljše rezultate dobimo v primeru neoptimalne izbire parametrov. To pa se ne zgodi v primeru fonetičnih značilk. Rezultati ostajajo konsistentni ne glede na izbiro uteži modelov. To je posledica dejstva, da so modeli govora in ne–govora naučeni na podlagi fonetičnih značilk veliko bolj diskriminatorni in je delovanje postopkov GNG segmentacije v tem primeru bolj stabilno.

Če primerjamo rezultate GNG segmentacije samo v primeru fonetičnih značilk, ne moremo ugotoviti kakšnih izrazitih posebnosti v delovanju različnih postopkov in različnih značilk. Tako s CVS kot VUS značilkami smo dobili dobre rezultate razpoznavanja. Tudi s postopki segmentacije nismo pridobili kakšnih izrazitih razlik v delovanju, čeprav smo v obeh primerih zbirk s segmentacijo *BICseg–GMM* dobili malce boljše rezultate. Lahko pa ugotovimo, da so predlagane fonetične značilke neodvisne od jezika razpoznavanja, saj smo v primeru obeh razpoznavalnikov glasov dobili zelo primerljive rezultate, ki se ujemajo s poteki rezultatov GNG segmentacije na razvojni zbirki.

Najboljše rezultate razpoznavanja smo dosegli s kombinacijo predstavitev v postopkih GNG segmentacije s fuzijo. V obeh primerih fuzije smo s kombinacijo dveh predstavitev presegli rezultate razpoznavanja samostojnih predstavitev. Prav tako kot v razvojni zbirki je bilo tudi tu v primeru fuzije *MFCC+ent.,din.* opazno izrazito odstopanje razpoznavanja govora in ne–govora v primeru optimalnih in neoptimalnih izbir uteži (še posebej v tabeli 3.3). Generalno gledano pa smo najboljše rezultate GNG segmentacije dosegli s fuzijo MFCC in CVS značilk. To govori v prid dejstvu, da z združevanjem akustične in fonetične informacije pridobimo dvoje: z akustičnimi značilkami povečamo detekcijo govora (predvsem kratkih segmentov), s fonetičnimi pa detekcijo ne–govornih pojavov, hkrati pa še povečamo stabilnost delovanja postopkov GNG segmentacije.

Če vse skupaj povzamemo, lahko na podlagi rezultatov iz razvojne zbirke in rezulta-

tov iz tabel 3.2 in 3.3 ugotovimo, da s predlaganimi fonetičnimi značilkami izboljšamo delovanje postopkov GNG segmentacije. To lahko razložimo z dejstvom, da so bile te značilke namensko pridobljene za detekcijo govornih in ne–govornih segmentov, medtem ko se značilke MFCC in značilke entropije–dinamizma uporabljajo širše. Poglavitna prednost fonetičnih značilk je predvsem v stabilnosti delovanja postopkov GNG segmentacije ob uporabi teh značilk. Medtem ko je zanesljivost delovanja postopkov z MFCC značilkami in značilkami entropije–dinamizma močno nihala, smo s CVS (VUS) značilkami dosegali podobne rezultate v različnih pogojih delovanja. Najboljše rezultate GNG segmentacije smo dosegli s kombinacijo MFCC in CVS značilk, s čimer smo pokazali, da z združevanjem akustične in fonetične informacije pridobimo najboljše predstavitve zvočnih posnetkov za GNG segmentacijo.

3.5 Zaključek

V tem poglavju smo se posvečali predstavitvam in postopkom segmentacije zvočnih posnetkov na govorne in ne–govorne odseke. V ta namen smo razvili novo predstavitev zvočnih signalov, s katero smo lahko izvajali detekcijo govora v segmentih samo na podlagi dveh modelov razvrščanja: modela govora in modela ne–govora. S tem smo sledili osnovnemu principu detekcije govora, kjer imamo definirana samo dva razreda razvrščanja in je razred ne–govora določen z razredom govora.

Predstavitev je temeljila na fonetični informaciji, ki smo jo pridobili iz samodejnih transkripcij govora na podlagi osnovnih razpoznavalnikov glasov. Na ta način smo izpeljali štiri osnovne mere značilk, ki so bile izvedene iz dveh kombinacij skupin glasov: parov samoglasnik – soglasnik in parov zvenečih in nezvenečih glasov. Značilke so bile načrtovane tako, da so bile neodvisne od jezika razpoznavanja in modelov osnovnih govornih enot, ki so bile vključene v razpoznavalnik. Pridobivanje značilk je bilo zasnovano tako, da smo jih lahko vključili v različne sisteme segmentacije posnetkov na govor in ne–govor. Preizkušali smo tudi dva postopka segmentacije. Oba sta temeljila na GMM modelih. V prvem postopku sta se izvajala segmentacija in razvrščanje istočasno. To smo dosegli s sestavljanjem GMM modelov v HMM modele. V alternativnem postopku segmentacije, ki smo ga razvili skupaj s fonetičnimi značilkami, pa sta potekala segmentacija in razvrščanje ločeno. V prvi fazi se je izvedla segmentacija posnetkov glede na akustične lastnosti signalov, nato pa smo izračunali fonetične značilke in sprožili postopek razvrščanja z GMM modeli.

Predlagane predstavitve in postopke smo primerjali z referenčnimi segmentacijami na dveh zbirkah zvočnih posnetkov informativnih oddaj, zbirki SiBN in zbirki COST278. Skupna ugotovitev vrednotenja postopkov GNG segmentacije na podlagi različnih predstavitev je bila, da je delovanje postopkov s fonetičnimi značilkami bolj robustno in stabilno ne glede na različne pogoje delovanja. Dodatna analiza rezultatov je pokazala, da so akustične predstavitve, ki smo jih v našem primeru modelirali z MFCC značilkami, in predstavitve na podlagi delovanja razpoznavalnikov, ki smo jih opisovali z entropijo in dinamizmom, zelo občutljive na spremenjene pogoje delovanja. Tako smo v vseh primerih teh predstavitev opazovali zelo spremenljive rezultate detekcije ne–govornih segmentov ob različnih izbirah uteži modelov detekcije. To pa se ni zgodilo v primeru

fonetičnih značilk, kjer se je skupno razpoznavanje govora in ne–govora na testnih zbirkah gibalo med 95% in 98% na glede na izbiro odprtih parametrov segmentacije. Na ta način smo lahko pokazali, da so fonetične značilke manj občutljive na različne akustične razmere in na različne ne–govorne situacije, ki jih lahko pričakujemo v zvočnih posnetkih.

Druga skupina preizkusov je zajemala postopke GNG segmentacije s kombinacijo različnih predstavitev zvočnih signalov. Osnovno vodilo je bilo, da bi z združevanjem osnovne - akustične informacije in višje - fonetične informacije izboljšali razpoznavanje krajših govornih ali ne–govornih odsekov, hkrati pa bi ohranili neobčutljivost segmentacije na različne pogoje delovanja. V primeru kombinacije fonetičnih in MFCC značilk nam je tako uspelo izboljšati rezultate GNG segmentacije v vseh preizkusih.

Najboljše predstavitve in postopke GNG segmentacije smo uporabili v nadaljevanju pri segmentaciji in razvrščanju segmentov po govorcih.

4 Samodejna segmentacija zvočnih posnetkov

V tem poglavju se bomo posvetili segmentaciji zvočnih posnetkov informativnih oddaj. Namen segmentacije je razdeliti eno ali večmodalne tokove podatkov v homogene dele - segmente - glede na določene lastnosti, ki jih vnaprej predpišemo. Tako se lahko izvaja segmentacija glede na govor/ne–govor, zamenjavo govorcev, spremembe akustičnega ozadja, spremembe kvalitete posnetka ipd.

V tem poglavju se bomo na primeru posnetkov informativnih oddaj omejili na segmentacijo zvočnih posnetkov glede na zamenjave govorcev in spremembe akustičnega ozadja.

V uvodu bomo natančneje formulirali problem segmentacije in pregledali nekaj temeljnih del s tega področja. V nadaljevanju bomo opisali dva obstoječa postopka segmentacije, ki smo jih preizkušali na naših zbirkah zvočnih posnetkov, in na podlagi katerih smo predlagali dva nova postopka segmentacije. Prvi predlagani postopek je izboljšana verzija osnovnega postopka segmentacije, le da je kriterij iskanja mej med segmenti določen z relativnim pragom. Drugi postopek pa temelji na združevanju dveh metod segmentacije na podlagi različne akustične informacije. Vse opisane postopke bomo primerjali med seboj na različnih zbirkah zvočnih posnetkov.

4.1 Uvod

Pri določanju segmentov glede na zamenjave govorcev (*ang. speaker change detection*) in/ali akustičnega ozadja (*ang. background change detection*) gre za iskanje časovnih mej, kjer pride do zamenjave govorca in/ali spremembe v akustičnem ozadju zvočnega posnetka. Segment je tako definiran kot del posnetka med dvema mejama, kjer se zahtevane lastnosti ne spreminjajo.

Postopki samodejne segmentacije zvočnih posnetkov glede na zamenjave govorcev *(sprememba po govorcih, SG)* in spremembe v akustičnem ozadju *(sprememba po govorcih in v akustičnem ozadju, SAG)* se uporabljajo v različnih sistemih govornih tehnologij. Običajno predstavljajo prvi korak pri obdelavi zvočnih posnetkov v sistemih, kjer je potrebno daljše posnetke 'smiselno' razdeliti na relativno kratke dele za nadaljnjo obdelavo. Takšnih sistemov je več, delimo pa jih na dve skupini. Prvo skupino tvorijo sistemi, ki temeljijo na samodejnem razpoznavanju govora, drugo pa sistemi, ki temeljijo na razpoznavanju govorcev. V primeru razpoznavanja govora s segmentacijo razdelimo zvočne posnetke na manjše dele, ki so primerni za razpoznavanje, s postopki SG ali SAG segmentacije pa skupaj s postopki rojenja pridobimo še informacijo o govorcih v takšnih posnetkih. V takem primeru lahko z uporabo tehnik prilagajanja modelov razpoznavanja glede na govorce znatno izboljšamo rezultate razpoznavanja [Zhang-02, Pusateri-02]. Podobno se v primeru razpoznavanja govorcev uporablja SG ali SAG segmentacija v postopkih sledenja in detekcije govorcev [Martin-00, Istrate-05, Moraru-05] ter indeksacije zvočnih posnetkov [Magrin-Chagnolleau-02] običajno v prvih fazah, kjer se izvaja strukturiranje posnetkov glede na govorce. V tem primeru z učinkovito SG ali SAG segmentacijo razdelimo posnetke na take dele, kjer en segment predstavlja enega govorca v nespremenjenih akustičnih pogojih. Zato govorimo o t.i. čistosti segmentov *(ang. segment purity)*. Učinkovita segmentacija je tista, pri kateri dosežemo visoko stopnjo čistosti segmentov *(ang. high segment purity)*.

Pri pregledu raziskovalnega področja postopkov segmentacije v uvodnem poglavju smo razvrstili postopke glede na namen uporabe in glede na metode uporabljene pri sami segmentaciji. Glede na metode uporabljene pri segmentaciji smo razdelili postopke na dve skupini: metode segmentacije s predhodnim učenjem modelov in metode s sprotnim odločanjem na podlagi mer podobnosti med segmenti. Pri SG in SAG segmentaciji se skoraj izključno uporabljajo metode iz druge skupine. Za to obstajata vsaj dva razloga, ki smo jih omenili že v uvodnem poglavju. Poglavitni razlog je ta, da se SG (SAG) segmentacija izvaja kot začetni postopek v sistemih obdelave zvočnih posnetkov in zato običajno nimamo na razpolago nobene informacije o stanju obdelovanih posnetkov (npr. koliko in kakšni govorci so v posnetku, kakšnega tipa akustična ozadja lahko pričakujemo ipd.) in zato ni možno pri uporabi predhodno naučenih modelov predvideti vseh možnih pričakovanih situacij. Drugi razlog pa je v tem, da se za detekcijo meje med dvema segmentoma ne moremo odločati samo na podlagi predhodnega znanja (predhodnih mej med segmenti), ampak vsakič znova na podlagi trenutne podobnosti ali različnosti med segmentoma. Zato je 'naraven' pristop reševanja problema SG (SAG) segmentacije uporaba mer podobnosti (različnosti) med dvema segmentoma. Pri takšni segmentaciji se tako predvsem ukvarjamo z izbiro kriterijev in predstavitvijo

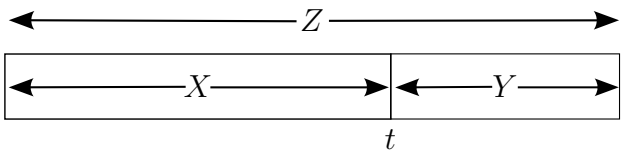
zvočnih signalov za določanje podobnosti (različnosti) med segmenti.

V nadaljevanju bomo tako v razdelku 4.2 najprej formulirali problem SG in SAG segmentacije in opisali dva referenčna postopka segmentacije, ki temeljita na Bayesovem informacijskem kriteriju *(ang. Bayesian information criterion, BIC)* kot meri podobnosti med dvema segmentoma. Bistvena lastnost takšnih in podobnih postopkov, kjer se uporabljajo mere podobnosti ali različnosti med segmentoma, je, da se sprejema odločitev o možni meji na podlagi vnaprej določenega praga združevanja. Tak prag se običajno oceni iz razvojne zbirke in ga je potrebno vedno znova prilagajati glede na akustične lastnosti posnetkov, ki jih obdelujemo. To predstavlja tudi glavno pomanjkljivost takšnih metod. V razdelku 4.4.1 je tako predstavljen postopek segmentacije z uporabo kriterija BIC, kjer se prag odločitve določa sprotno na podlagi zvočnega posnetka, ki ga trenutno obdelujemo, in tako ni že vnaprej podan. Takšen pristop k segmentaciji nam je omogočil tudi normalizacijo ocen odločitev *(ang. score normalization)*, zato smo lahko razvili še drugi postopek segmentacije z združevanjem BIC ocen na podlagi različnih značilk segmentov, ki je predstavljen v razdelku 4.4.2. Predlagane metode segmentacije smo primerjali z referenčnima metodama na zbirkah SiBN in COST278. Rezultati segmentacije so predstavljeni v razdelku 4.5.

4.2 Formulacija problema segmentacije

V postopkih SG ali SAG segmentacije, ki temeljijo na merah podobnosti (različnosti) med segmenti, izvedemo iskanje mej na podlagi odločitve, ali sta dva sosedna odseka analiziranega zvočnega signala med seboj podobna (različna) ali ne.

Če označimo levi odsek zvočnega posnetka z X, desnega pa z Y, kot je to prikazano na sliki 4.1, potem je kandidat za mejo med njima tista točka t, kjer v primeru mere podobnosti med odsekoma dosežemo najmanjšo vrednost kriterijske funkcije oziroma v primeru mere različnosti največjo vrednost. Odločitev, ali je točka t meja ali ne, pa običajno sprejmemo na podlagi vnaprej podanega praga segmentacije.



Slika 4.1: Odseka X in Y zvočnega signala, kjer se odločamo ali postavimo mejo t ali ne.

Učinkovito iskanje mej med odseki zvočnih posnetkov v primeru SG (SAG) segmentacije je torej odvisno od pravilne izbire mere podobnosti (različnosti) med odseki, postavitve praga odločitve za detekcijo meje in same predstavitve odsekov zvočnih signalov. Za segmentacijo celotnega zvočnega posnetka pa je bistvenega pomena postopek iskanja segmentov. V okviru doktorskega dela smo se tako ukvarjali s pravilno postavitvijo pragov za detekcijo mej, s predstavitvami zvočnih signalov in s postopki segmentacije.

Za mere podobnosti ali različnosti med odseki se uporabljajo predvsem mere verjetnostne različnosti [Pavešić-00, str. 168] in informacijski kriteriji [Burnham-03]. V prvem primeru primerjamo dva sosedna odseka na podlagi verjetnostnih porazdelitev ocenjenih iz predstavitev zvočnih signalov danih odsekov, v drugem pa primerjamo modele naučene iz danih odsekov. Primerjava dveh sosednih odsekov je lahko neposredna (uporabimo mero primerljivosti neposredno na odsekih X in Y iz slike 4.1), lahko pa jih primerjamo posredno preko skupnega odseka (primerjamo odseka X in Y s skupnim odsekom Z na sliki 4.1). Med najbolj uveljavljenimi merami verjetnostne različnosti pri SG (SAG) segmentaciji je simetrična Kullback–Leiblerjeva mera KL2 [Pavešić-00, str. 168], pri informacijskih kriterijih pa je najbolj uveljavljen kriterij BIC [Scwartz-76, Fraley-98].

V različnih študijah [Cettolo-00, Cettolo-05, Žibert-04] je bilo pokazano, da so najboljše mere za SG (SAG) segmentacijo tiste, pri katerih se izvaja primerjava odsekov posredno preko skupnega odseka, in med njimi je bil najboljši kriterij BIC. Zato smo za izbrano mero podobnosti v vseh naših postopkih uporabljali kriterij BIC, ki je opisan v naslednjem razdelku.

4.2.1 Kriterij BIC

Kriterij BIC je prvi predlagal Schwartz [Scwartz-76], pri SG (SAG) segmentaciji pa sta ga prva uporabila Chen in Gopalakrishnan [Chen-98], ki sta tudi prva formulirala problem SG segmentacije kot problem izbire pravih modelov za opisovanje odsekov zvočnih posnetkov. Osnovna lastnost kriterija BIC je namreč v tem, da se za vsaka dva sosedna odseka sprašujemo, ali jih je boljše opisati z dvema ločenima modeloma ali z enim skupnim modelom.

Formulirajmo problem bolj natančno. Denimo, da segmenta X in Y iz slike 4.1 opišemo z zaporedjem vzorcev $X = \{x_1, x_2, \ldots, x_{N_x}\}$ in $Y = \{y_1, y_2, \ldots, y_{N_y}\}$, kjer sta N_x in N_y števili vzorcev v obeh odsekih. Označimo skupni odsek Z kot unijo obeh odsekov, torej $Z = X \cup Y$, s skupnim številom vzorcev $N = N_x + N_y$. Pri kriteriju BIC za model predstavitve odsekov izberemo funkcijo porazdelitve gostote verjetnosti $p(. \mid \theta)$, kjer θ predstavlja parametre porazdelitve p. V tem primeru predpostavljamo, da so vzorci $\{x_i\}$ in $\{y_i\}$ predstavljeni z naključnimi spremenljivkami, ki so enako porazdeljene in med seboj neodvisne (*ang. independent and identically distributed, IID*). Označimo parametre porazdelitve odsekov X, Y in Z s θ_X, θ_Y in θ_Z. Običajno se za p izbere funkcijo gostote verjetnosti normalne porazdelitve, tako da so parametri takšnega modela predstavljeni z ocenjenim povprečnim vektorjem $\mathbf{m}$ in kovariančno matriko $\mathbf{\Sigma}$ vzorcev danega odseka.

Iskanje meje med dvema odsekoma pri SG (SAG) segmentaciji se tako pri kriteriju BIC prevede na odločitev med dvema hipotezama. Predpostavka prve hipoteze H_0 je, da v času t ni meje med odsekoma X in Y oziroma, da je zaporedje vzorcev $\{x_i\}$ in $\{y_i\}$ odsekov X in Y prispeval isti govorec pri SG segmentaciji oziroma isti vir pri SAG segmentaciji. V tem primeru lahko logaritem vrednosti gostote verjetnosti *(ang.*

log-likelihood, LLH) za hipotezo H_0 zapišemo kot:

$$L_0 = \sum_{n=1}^{N_x} \log p(x_n \mid \theta_Z) + \sum_{n=1}^{N_y} \log p(y_n \mid \theta_Z). \tag{4.1}$$

Pri drugi hipotezi H_1 je predpostavka, da točka v času t predstavlja mejo med odsekoma X in Y oziroma, da sta zaporedje vzorcev odseka X in Y v primeru SG segmentacije tvorila dva različna govorca (v primeru SAG segmentacije dva različna vira). V tem primeru je logaritem vrednosti gostote verjetnosti (LLH) enak:

$$L_1 = \sum_{n=1}^{N_x} \log p(x_n \mid \theta_X) + \sum_{n=1}^{N_y} \log p(y_n \mid \theta_Y). \tag{4.2}$$

Kriterij BIC v primeru SG (SAG) segmentacije je tako definiran kot [Ajmera-04]:

$$d_{BIC} = L_1 - L_0 - \frac{\lambda}{2} \cdot \Delta K \cdot \log N, \tag{4.3}$$

kjer sta L_0 in L_1 definirana z enačbama (4.1) in (4.2), N je število vzorcev v skupnem odseku Z, ΔK predstavlja razliko v številu parametrov med modeloma iz hipoteze H_1 in H_0, λ pa je utežni faktor kriterija BIC. V primeru Gaussovih porazdelitev se kriterij prevede na [Chen-98]:

$$d_{BIC} = \frac{N}{2} \log |\mathbf{\Sigma}_Z| - \frac{N_x}{2} \log |\mathbf{\Sigma}_X| - \frac{N_y}{2} \log |\mathbf{\Sigma}_Y| - \frac{\lambda}{2}(d + \frac{d(d-1)}{2}) \cdot \log N, \tag{4.4}$$

kjer so $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_Y$ in $\mathbf{\Sigma}_Z$ ocenjene kovariančne matrike odsekov X, Y in Z ter d dimenzija vektorjev vzorcev danih segmentov.

Razlika $L_1 - L_0$ v enačbi (4.3) predstavlja razliko v oceni podatkov iz modelov hipotez H_1 in H_0, drugi del enačbe $\Delta K \cdot \log N$ pa razliko v kompleksnosti modelov. Razlika $L_1 - L_0$ je vedno pozitivna, saj podatke iz odseka Z v primeru L_1 ocenjujemo z dvema porazdelitvama in je zato LLH večji (model je boljši), v primeru L_0 pa jih ocenjujemo samo z eno porazdelitvijo in je zato LLH manjši. Po drugi strani pa imamo v primeru dveh modelov (dveh porazdelitev) še enkrat več parametrov kot v primeru enega modela (porazdelitve) in je zato kompleksnost modela pri hipotezi H_1 večja kot v primeru H_0. Z utežnim faktorjem λ tako uravnavamo razmerje med kvaliteto ocen in kompleksnostjo modelov. V splošnem velja, večja kot je vrednost d_{BIC}, boljše opišemo dane podatke s kompleksnejšim modelom (torej z dvema porazdelitvama), manjša kot je vrednost d_{BIC}, boljše opišemo podatke z manj kompleksnim modelom (torej z eno porazdelitvijo). To v primeru SG (SAG) segmentacije pomeni, večja kot je vrednost d_{BIC}, bolj verjetna je hipoteza H_1, in obratno, manjša kot je vrednost d_{BIC}, bolj verjetno velja hipoteza H_0. Odločitev med obema hipotezama se uravnava z utežnim faktorjem λ in velja, če je $d_{BIC} > 0$, velja hipoteza H_1, če je $d_{BIC} < 0$, pa velja hipoteza H_0.

V osnovni definiciji kriterija BIC [Scwartz-76] je $\lambda = 1$, vendar se je v primeru SG (SAG) segmentacije izkazalo, da je potrebno za doseganje optimalnih rezultatov spreminjati vrednost λ. Tako je Kemp s sod. [Kemp-00] pokazal, da so rezultati segmentacije močno odvisni od spreminjanja vrednosti λ in so pri $\lambda = 1$ slabši,

kot pa pri izbiri optimalnih vrednosti. Številni avtorji uporabljajo različne vrednosti λ, [Tritschler-99, Vandecatseye-03], ki jih določajo izkustveno ali na podlagi razvojnih zbirk [Delacourt-01]. V naših preizkusih smo določali vrednosti λ glede na optimalne rezultate segmentacije na razvojni zbirki.

λ predstavlja implicitno določen prag segmentacije, pri katerem se odločamo za eno izmed hipotez H_0 ali H_1. Kot se je pokazalo v številnih eksperimentih segmentacije [Chen-02, Ajmera-04, Vandecatseye-03] in smo opazili tudi mi pri naših preizkusih, z večjimi vrednostmi λ ne uspemo detektirati vseh mej med dejanskimi segmenti, z manjšimi vrednostmi λ pa detektiramo preveč mej, kot je dejanskih.

4.3 Referenčne metode segmentacije

V nadaljevanju bomo opisali dva referenčna postopka segmentacije, ki delujeta na podlagi kriterija BIC. Predstavljena referenčna postopka se razlikujeta v načinu iskanja kandidatov za meje med posameznimi segmenti, medtem ko za odločitev, ali je predlagani kandidat res meja, uporabljata kriterij BIC. Predstavljena sta oba postopka, ker smo na podlagi obeh pristopov zgradili nov postopek segmentacije z normalizacijo ocen odločitev za meje, ki ga bomo opisali v naslednjem razdelku.

4.3.1 Osnovni postopek segmentacije s kriterijem BIC

Postopek SG (SAG) segmentacije zvočnih posnetkov, ki bo opisan v nadaljevanju, sta predlagala Chen in Gopalakrishnan [Chen-98].

Osnovna ideja postopka je, da se začetni osnovni odsek, kjer iščemo potencialna segmenta, povečuje toliko časa, dokler s kriterijem BIC ne najdemo meje med levim in desnim delom danega odseka. Ko takšno mejo najdemo, jo označimo s t, postavimo odsek iskanja meje na osnovno dolžino z začetkom v točki $t+1$ in ponovimo postopek iskanja. Postopek segmentacije [Ajmera-04, Chen-98] se tako izvaja po naslednjih korakih:

1. *določi interval iskanja meje* $[a, b]$
   $a = 0$; $b =$ `MIN_ODSEK`;

2. *poišči kandidata za mejo na intervalu [a, b] glede na kriterij BIC*
   $t = \arg\max_{[a,b]} d_{BIC}$;

3. *če je* $d_{BIC} < 0$ *v točki t, potem velja hipoteza* H_0
   $b = b +$ `DODANI_VZORCI`;

   če je $d_{BIC} > 0$ *v točki t, potem velja hipoteza* H_1
   $a = t + 1$; $b = a +$ `MIN_ODSEK`;

4. *če velja* $b - a >$ `MAX_ODSEK`, *potem*
   $a = b -$ `MAX_ODSEK`; $b = a +$ `MIN_ODSEK`;

5. *ponovi točko 2.*

Poleg λ v kriteriju BIC, d_{BIC}, so odprti parametri postopka še `MIN_ODSEK`, `DODANI_VZORCI` in `MAX_ODSEK`, ki jih običajno določimo iz razvojnih zbirk glede na optimalne rezultate segmentacije in glede na željeno hitrost delovanja algoritma. Dejstvo je namreč, da za vsak izračun kriterija d_{BIC} potrebujemo nove ocene modelov levega, desnega in skupnega odseka na intervalu $[a, b]$ in večji kot je odsek, več časa potrebujemo za izračun parametrov modelov. Pri uporabi Gaussovih porazdelitev za modele segmentov je potrebno pri iskanju največje vrednosti kriterija BIC iz (4.4) za vse t iz intervala $[a, b]$ vsakič znova ocenjevati kovariančne matrike levega (X) in desnega odseka (Y), medtem ko se za skupni odsek (Z) izračuna kovariančna matrika samo enkrat na iteracijo. Zato so bile v [Tritschler-99] in v [Cettolo-05] predlagane številne izboljšave za pohitritev postopka segmentacije, predvsem kriterija BIC, ki jih omogoča uporaba Gaussovih porazdelitev za modeliranje segmentov.

Opozoriti velja tudi, da se pri iskanju kandidatov za možno mejo med segmenti izbere samo tisti t, pri katerem doseže vrednost d_{BIC} lokalni maksimum na intervalu $[a, b]$ (korak 2 v postopku), kljub temu da lahko za več t-jev iz intervala velja hipoteza H_1 (torej $d_{BIC} > 0$ pri danem t). Izbiro takšnega t, pri katerem doseže d_{BIC} maksimalno vrednost, lahko utemeljimo z dejstvom iz razdelka 4.2.1, da večja kot je vrednost d_{BIC}, bolj verjetna je hipoteza H_1.

Opisani postopek segmentacije z izboljšavami, predlaganimi v [Tritschler-99], smo uporabili kot referenčno metodo pri vrednotenju postopkov segmentacije v naših preizkusih.

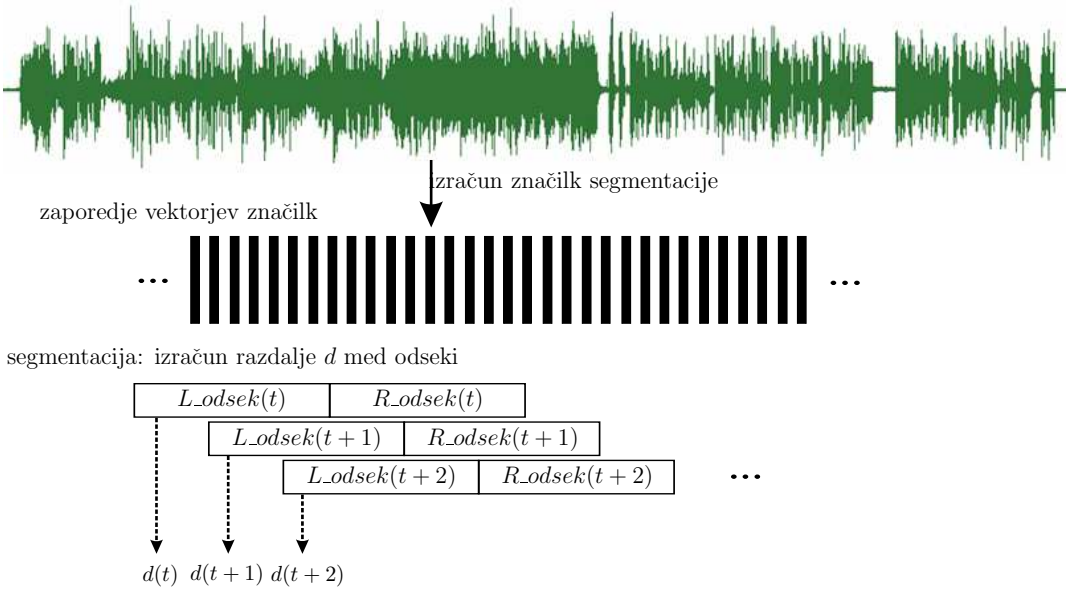
4.3.2 Postopek segmentacije DISTBIC

Postopek segmentacije DISTBIC je bil predlagan v [Delacourt-01] in je v bistvu izboljšana verzija postopka [Siegler-97], ki so ga uporabljali kot referenčni postopek segmentacije v evaluacijah *NIST Hub-4* samodejnega podnaslavljanja informativnih oddaj [Pallett-02].

Postopek segmentacije DISTBIC poteka v dveh fazah. V prvi poiščemo kandidate za meje med segmenti, v drugi pa se na podlagi kriterija BIC odločamo, ali je kandidat res meja ali ne. Tako se v postopku DISTBIC predpostavlja, da imamo celotne posnetke že na razpolago, medtem ko je pri osnovnem postopku BIC možna sprotna segmentacija posnetkov.

Na sliki 4.2 je prikazana prva faza segmentacije DISTBIC. Tu se odločamo o kandidatih za možne meje med segmenti na podlagi mere razdalje[1] med odseki (od tod tudi prvi del imena metode, DIST je okrajšava za razdaljo *(ang. distance)*). Postopek iskanja kandidatov za mejo se v tem primeru prevede na iskanje lokalnih maksimumov kriterijske funkcije, ki je definirana kot mera razdalje med dvema sosednima enako dolgima odsekoma zvočnega signala. Tako se razdalja $d(t)$ v točki t iz slike 4.2 izračuna kot vrednost mere podobnosti med odsekoma $L\_odsek(t)$ in $R\_odsek(t)$. Odseki L_odsek in R_odsek so enakih dolžin za vse t. Dolžine odsekov izbiramo tako dolge,

[1] Za mero razdalje se ne uporablja nujno *metrika (razdalja)* v matematičnem smislu, ampak se lahko uporablja poljubna mera podobnosti (različnosti) med segmentoma. Zato uporabljamo izraz *mera razdalje* in ne samo razdalja.

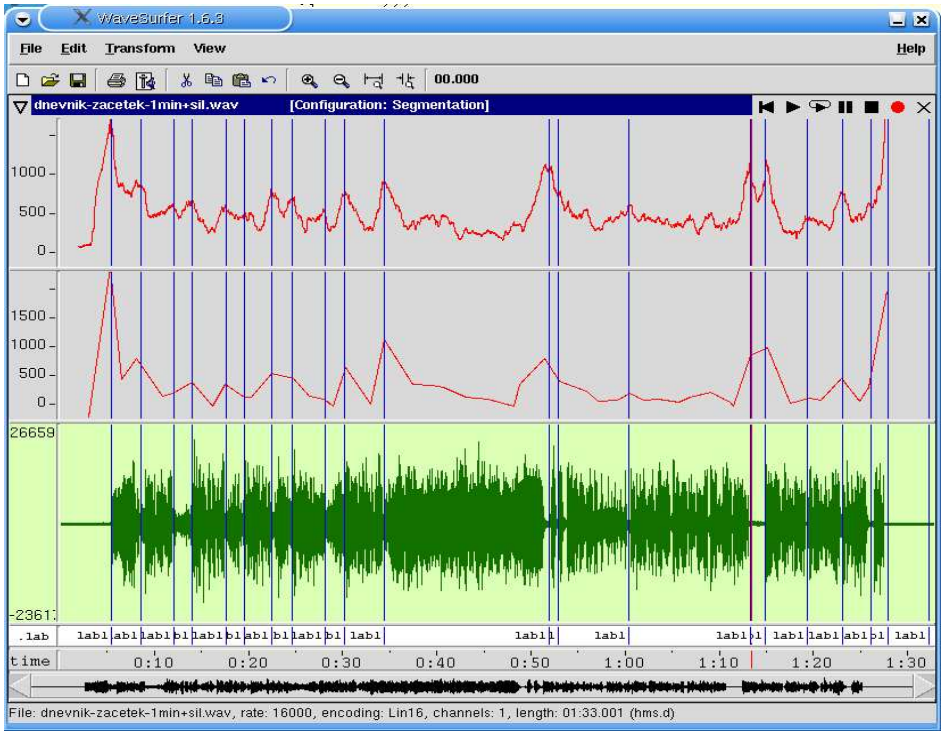


Slika 4.2: Prva faza segmentacije DISTBIC. Izračun razdalj d na enako dolgih levih in desnih odsekih za vsak t po celotnem posnetku.

da lahko dovolj dobro ocenimo vrednosti mere podobnosti, ki jo računamo. Običajne dolžine odsekov za segmentacijo so okoli 2.0 s, [Delacourt-01]. Razdalja v času $t+1$ se izračuna povsem enako kot za točko t, le da se premaknemo v signalu za predpisani premik naprej (lahko je en vzorec - en vektor v zaporedju vektorjev značilk, s katerimi predstavimo analizirani signal, ponavadi to pomeni, da se premikamo za nekaj deset milisekund analiziranega signala) in ponovimo izračun za vrednost razdalje d na premaknjenih odsekih. Ko izračunamo razdalje d za vse predpisane t, poiščemo med njimi točke, kjer vrednosti kriterijske funkcije dosežejo lokalne maksimume. Kandidat za mejo postane tisti lokalni maksimum zglajene verzije kriterijske funkcije, ki ustreza določenim dodatnim pogojem, predpisanim v [Delacourt-01].

Druga faza segmentacije DISTBIC poišče dejanske meje med kandidati za mejo iz prve faze postopka. Tu se za detekcijo meje uporablja kriterij BIC (od tod tudi drugi del imena metode). Meja med segmentoma postane tista točka t, za katero velja hipoteza $H1$, torej $d_{BIC} > 0$ v točki t. Točke, za katere je $d_{BIC} < 0$, izločimo in segmenta združimo.

V obeh fazah postopka nastopajo mere podobnosti (različnosti) med segmenti. V osnovni verziji postopka [Siegler-97] so tako v obeh fazah uporabljali simetrično Kullback-Leiblerjevo mero, medtem ko so v [Delacourt-01] testirali različne mere za iskanje kandidatov za mejo v prvi fazi, v drugi fazi pa so uporabljali kriterij BIC. Najboljše rezultate segmentacije so dobili, ko so za kriterijsko funkcijo izbrali razliko v LLH-jih skupnega odseka in vsote levega in desnega odseka, $L_1 - L_0$ iz enačbe (4.3), *(ang. generalized log-likelihood ratio, GLLR)*. Tudi mi smo zato v vseh naših preizkusih segmentacije s postopkom DISTBIC uporabili v obeh fazah kriterij BIC, pri čemer smo v prvi fazi λ postavili na 0.0, v drugi fazi postopka pa smo uporabljali optimalne λ pridobljene iz razvojnih zbirk.



Slika 4.3: Postopek DISTBIC segmentacije na delu posnetka informativne oddaje. V zgornjem oknu je prikazan potek kriterijske funkcije na podlagi d_{BIC} iz prve faze postopka. V srednjem oknu so prikazane izračunane vrednosti d_{BIC} iz druge faze na kandidatih za mejo, ki smo jih določili v prvi fazi postopka. V spodnjem oknu je prikazan zvočni signal skupaj z dejanskimi mejami segmentov različnih govorcev, ki so predstavljene z navpičnimi črtami po celotni sliki.

Na sliki 4.3 je prikazan primer segmentacije s postopkom DISTBIC. V zgornjem oknu je prikazan potek kriterijske funkcije GLLR v primeru dela posnetka informativne oddaje. V srednjem oknu so prikazane vrednosti kriterija BIC iz druge faze pri $\lambda = 1.0$ v točkah lokalnih maksimumov iz prve faze postopka. Spodnje okno pa prikazuje zvočni signal skupaj z dejanskimi mejami med segmenti. Kot lahko vidimo, se v večini primerov lokalni maksimumi kriterijske funkcije ujemajo z dejanskimi mejami med segmenti. Tudi vrednosti d_{BIC} pri $\lambda = 1.0$ v srednjem oknu potrjujejo, da so izbrane točke res meje med segmenti, težava pa je v tem, da so lokalni maksimumi v nekaterih primerih zelo izraziti, v nekaterih primerih pa ne. Zato je potrebno izvajati izredno natančno iskanje lokalnih maksimumov, ki predstavljajo prave kandidate za meje. To pa predstavlja tudi glavno pomanjkljivost te metode.

4.4 Predlagane metode segmentacije

V prejšnjih razdelkih smo si ogledali dva postopka segmentacije, ki temeljita na kriteriju BIC za določitev meje med segmenti.

Kriterij BIC se je v različnih preizkusih segmentacij izkazal za izredno učinkovito mero podobnosti med segmenti. To pa predvsem zato, ker se primerjava med segmenti ne

izvaja neposredno na množicah vzorcev analiziranih segmentov, ampak posredno preko modelov predstavitev danih segmentov. S tem se doseže manjšo občutljivost na spremembe v zvočnih signalih in se obravnava segmente kot enovite celote. Druga prednost kriterija BIC s podobnimi merami je tudi v tem, da se dva segmenta ne primerjata neposredno med seboj, ampak da je izvedena primerjava preko skupnega segmenta. Na ta način se izognemo absolutnim ocenam primerjave in dejansko merimo relativne odmike, koliko se levi in desni segment, predstavljena z ločenima modeloma, razlikujeta od skupnega segmenta, predstavljenega z enim modelom. Kriteriji, ki delujejo na podoben način (informacijski kriteriji AIC[2], MDL[3], ...), so bili tudi že preizkušeni v postopkih segmentacije [Cettolo-00, Žibert-04], vendar se je z njimi dosegalo slabše ali primerljive rezultate s kriterijem BIC. Zato se je kriterij BIC uveljavil kot skoraj edini kriterij za SG (SAG) segmentacijo.

Glavna pomanjkljivost postopkov segmentacije, ki temeljijo na kriteriju BIC, je v odločitvi, ali je dana točka meja med segmentoma ali ne. Ta odločitev se sprejema na podlagi vrednosti d_{BIC} in pogoja $d_{BIC} \lessgtr 0$. Ali je vrednost d_{BIC} večja ali manjša od 0, pa je odvisno od izbire λ, ki tako implicitno določa prag odločitve. Izbira prave λ je odvisna tako od predstavitve zvočnih signalov, ki jih obdelujemo, kot od postopka segmentacije in nastavitve odprtih parametrov postopkov (npr. `MIN_ODSEK`, `DODANI_VZORCI` in `MAX_ODSEK` pri osnovni metodi, ali pa dolžina odsekov L_odsek in R_odsek pri metodi DISTBIC). Prag odločitve je potrebno prilagajati zaradi različne kvalitete in tipov zvočnih posnetkov, ki jih obdelujemo, zato je najpogostejši način določitve λ in ostalih odprtih parametrov postopkov segmentacije na podlagi optimalnih rezultatov segmentacije na razvojnih zbirkah [Delacourt-01, Vandecatseye-03]. Drugi način je izbira kriterijev in predstavitev zvočnih signalov, ki so manj občutljivi na pričakovane spremembe v kvaliteti in tipu posnetkov, [Ajmera-03].

V okviru doktorskega dela bomo v nadaljevanju predstavili drugačno rešitev problema iskanja optimalnega praga odločitve za mejo. Ker se prag odločitve spreminja od posnetka do posnetka in je močno odvisen tudi od drugih odprtih parametrov postopka segmentacije, ki jih lahko sprotno spreminjamo, predlagana metoda temelji na relativno določenem pragu, ki ga sprotno prilagajamo glede na dani posnetek. V ta namen smo združili oba referenčna postopka segmentacije: postopek DISTBIC smo uporabili za oceno praga odločitve, osnovni postopek, opisan v razdelku 4.3.1, s prilagodljivim pragom odločitve pa za detekcijo mej med segmenti.

4.4.1 Postopek segmentacije s kriterijem BIC in relativno določenim pragom

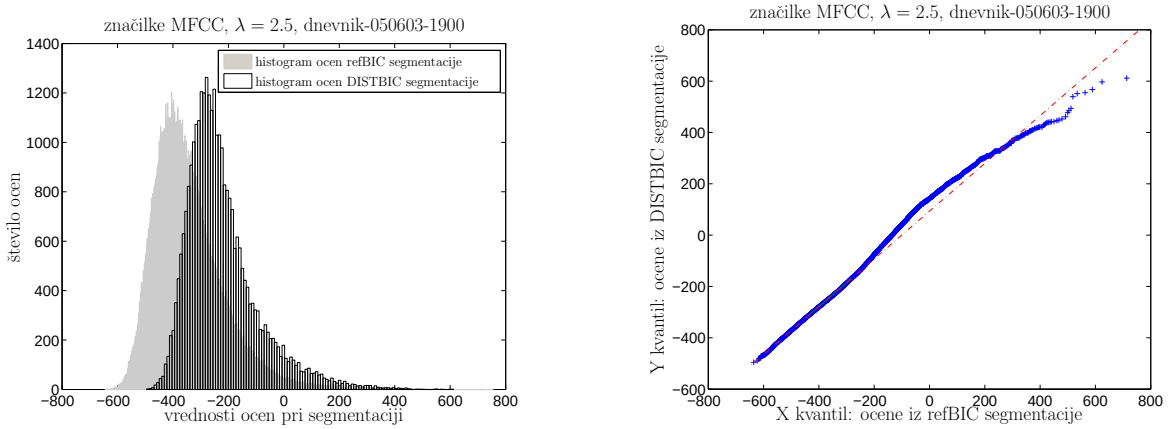
Ker so se postopki segmentacije s kriterijem BIC izkazali za učinkovite v primeru optimalne izbire prostih parametrov in konstantnih pogojev zvočnih posnetkov, smo se tudi mi v okviru doktorskega dela odločili za SG (SAG) segmentacijo z uporabo kriterija BIC. Kot je bilo že nakazano v prejšnjem razdelku, največji problem takšnih postopkov predstavlja prav izbira optimalnih odprtih parametrov postopkov, predvsem izbira

[2]AIC je kratica za Akaike Information Criteria.
[3]MDL je kratica za mero Minimum Description Length.

praga odločitve za mejo, ki ga predstavlja λ. V nadaljevanju bomo opisali postopek segmentacije, kjer prag odločitve za mejo sprotno prilagajamo glede na spremembe zvočnih razmer v posnetkih, ki jih obdelujemo.

V predlaganem postopku smo združili oba postopka segmentacije: osnovni postopek (*refBIC*), opisan v razdelku 4.3.1, in postopek *DISTBIC*. Oba postopka temeljita na kriteriju BIC, pa tudi segmentacija se izvaja na podoben način. Pri postopku *DISTBIC* najprej na podlagi lokalnih maksimumov poiščemo kandidate za meje v prvi fazi, v drugi pa jih izločamo, pri postopku *refBIC* pa obe fazi opravimo hkrati. Če tudi v prvi fazi postopka *DISTBIC* za iskanje kandidatov za meje uporabimo kriterij BIC, postane delovanje obeh postopkov zelo podobno. V obeh primerih se v točkah maksimumov kriterija BIC odločamo za mejo med segmentoma, razlika je le v tem, da v primeru postopka *DISTBIC* iščemo lokalne maksimume na celotnem posnetku, v primeru *refBIC* pa iščemo globalne maksimume na krajših odsekih danega posnetka (na intervalu $[a, b]$ iz koraka 2 postopka *refBIC*). V obeh primerih so torej kandidati za meje lokalni maksimumi, le da jih v primeru postopka *DISTBIC* iščemo globalno na podlagi vrednosti kriterijske funkcije na fiksno določenih odsekih izračuna kriterija BIC, medtem ko v primeru postopka *refBIC* iščemo lokalno med vrednostmi BIC izračunanimi na spremenljivih odsekih. Na splošno torej velja, da pri postopku *DISTBIC* v primeru, če za mero podobnosti uporabimo kriterij BIC, računamo vrednosti BIC na vedno enako določenih odsekih, pri postopku *refBIC* pa na spremenljivih odsekih. To pa pomeni, da se vrednosti ocen kriterija BIC v obeh primerih približno enako obnašajo. Se pravi, da v okolicah potencialnih mej med segmenti dobimo višje vrednosti kriterija BIC, na področjih, kjer ni meje pa primerjalno nižje vrednosti ne glede na izbrano λ.

Primerjava vrednosti ocen kriterija BIC pri segmentaciji z refBIC in DISTBIC



(a) (b)

Slika 4.4: Primerjava vrednosti ocen kriterija BIC pri segmentaciji s postopkoma *refBIC* in *DISTBIC* v primeru ene ure posnetka TV dnevnika. Slika (a) prikazuje histograma vrednosti ocen kriterija BIC obeh segmentacij, na sliki (b) pa je graf kvantil–kvantil primerjav.

To dejstvo je lepo razvidno iz prikazov na sliki 4.4. Tu smo primerjali vrednosti kriterija BIC v postopku *refBIC* in *DISTBIC* segmentacije na približno eni uri posnetka informativne oddaje. V obeh primerih smo izvajali segmentacijo na podlagi standardnih statičnih značilk koeficientov melodičnega kepstra MFCC pri $\lambda = 2.5$. Na sliki 4.4 (a)

je prikazan histogram porazdelitev ocen kriterija BIC v primeru obeh postopkov. Kot lahko ugotovimo, se ocenjene porazdelitve vrednosti obeh segmentacij ujemajo, le da sta malce zamaknjeni. To je posledica dejstva, da smo pri *DISTBIC* segmentaciji uporabili okno izračuna kriterija BIC dolžine 3.0 s, medtem ko pri *refBIC* segmentaciji računamo kriterij BIC, d_{BIC}, na spremenljivih odsekih dolžine od 2.0 do 10.0 s. Ujemanje porazdelitev ocen kriterija BIC je še bolj razvidno na sliki 4.4 (b), kjer smo primerjali med seboj kvantile obeh porazdelitev. Prekinjena premica na sliki 4.4 (b) predstavlja ekstrapolirano premico med prvim in tretjim kvartilom obeh porazdelitev in služi kot merilo za linearno ujemanje obeh porazdelitev. V splošnem velja, bližje kot so točke kvantilov obeh porazdelitev ekstrapolirani premici, večja je linearna odvisnost obeh porazdelitev. V našem primeru je ujemanje veliko, zato lahko zaključimo, da se ocene kriterijev BIC v obeh segmentacijah porazdeljujejo enako.

To spoznanje nam je omogočilo, da smo lahko iz postopka segmentacije *DISTBIC* segmentacije ocenjevali vrednosti kriterijske funkcije postopka *refBIC*. V našem primeru nas je zanimal predvsem interval porazdelitve *DISTBIC* ocen, saj smo na podlagi minimalne in maksimalne vrednosti ocen postopka *DISTBIC* lahko sklepali na interval porazdelitve ocen iz postopka *refBIC*. Tako smo na podlagi intervala minimalne in maksimalne vrednosti *DISTBIC* ocen določili prag odločitve za mejo v postopku *refBIC*.

Postopek je potekal v dveh fazah. V prvi fazi smo izračunali vrednosti kriterija BIC na vnaprej določenih odsekih analize danega posnetka. V bistvu smo izvedli prvo fazo postopka *DISTBIC*. Iz izračunanih ocen smo poiskali minimalno ($minBIC_{DISTBIC}$) in maksimalno vrednost ocen ($maxBIC_{DISTBIC}$) ter tako določili interval porazdelitev vrednosti *DISTBIC*. Na podlagi relativnega deleža ocen α, ki je bil podan kot vhodni parameter predlaganega postopka segmentacije, smo določili prag segmentacije postopka *refBIC* po formuli:

$$\theta = maxBIC_{DISTBIC} - \alpha \cdot (maxBIC_{DISTBIC} - minBIC_{DISTBIC}). \qquad (4.5)$$

Tako smo določili relativni prag segmentacije postopka *refBIC*. Prag θ je posredno odvisen od izbire λ in se od posnetka do posnetka razlikuje glede na ocenjene vrednosti kriterija BIC v postopku *DISTBIC*. Relativni delež ocen α smo izbirali med 0.0 in 0.2.

Druga faza postopka je bila segmentacija s postopkom *refBIC*, opisanem v razdelku 4.3.1. Razlika je bila le v tem, da smo prag odločitve za mejo v koraku 3 postopka *refBIC* postavili na θ, torej če je bil $d_{BIC} > \theta$ v točki t, potem je bila točka t meja, če pa je veljalo $d_{BIC} < \theta$, potem točka t ni bila meja. Pri tem smo seveda predpostavili, da sta λ v postopku *DISTBIC* in *refBIC* enaki.

Z združitvijo obeh postopkov smo tako v predlaganem postopku segmentacije odpravili potrebo po iskanju optimalnih vrednosti λ. Hkrati smo na podlagi relativnega deleža ocen α definirali relativni prag odločitve za meje med segmenti, ki se spreminja glede na zvočne razmere v obdelovanih posnetkih. Relativni delež α predstavlja tudi bolj naravno izbiro praga kot pa absolutne vrednosti λ, saj z njim določimo, kolikšen delež maksimalne ocene naj predstavlja prag odločitve.

Omenimo še, da bi idejo o uvedbi relativnega praga lahko izvedli tudi v postopku *DISTBIC* pri iskanju lokalnih maksimumov za kandidate za meje. Med njimi bi lahko izbirali samo tiste lokalne maksimume, ki ležijo nad pragom θ. Izkazalo pa se je, da

je ravno iskanje lokalnih maksimumov najbolj problematičen del postopka *DISTBIC*, zato smo relativni prag θ raje uporabili pri segmentaciji s postopkom *refBIC*.

4.4.2 Postopek segmentacije z združevanjem različnih predstavitev zvočnih posnetkov

Drugi stranski učinek ocenjevanja vrednosti referenčnega postopka *refBIC* iz vrednosti postopka *DISTBIC* je tudi možna normalizacija ocen kriterija BIC v osnovnem postopku. To nam je omogočilo uravnoteženo združevanje ocen segmentacije iz osnovnega postopka na podlagi različnih predstavitev zvočnih signalov. V nadaljevanju bomo zato opisali postopek, ki temelji na združevanju ocen kriterija BIC iz različnih predstavitev na podlagi *min–max normalizacije* vrednosti ocen [Jain-05].

Predlagani postopek predstavlja segmentacijo na podlagi odločitev, sprejetih iz skupnih normaliziranih in uravnoteženih ocen odločitev kriterija BIC iz več posamičnih segmentacij, zato govorimo o t.i. segmentaciji na podlagi fuzije ocen kriterija BIC (*ang. score-based fusion*). Osnovna ideja predlagane segmentacije je bila v tem, da bi združevali ocene kriterija BIC dveh različnih predstavitev zvočnih posnetkov z namenom povečanja zanesljivost odločanja o mejah med segmenti.

Postopek z združevanjem ocen kriterija BIC (*fuzBIC*) je tako potekal v dveh fazah. V prvi fazi smo najprej ocenili vrednosti kriterija BIC na podlagi postopka *DISTBIC*, kot je bilo že opisano v prejšnjem razdelku. Tako pridobimo ocene za minimalno vrednost kriterija BIC, $minBIC^{S_i}_{DISTBIC}$, in ocene za maksimalno vrednost, $maxBIC^{S_i}_{DISTBIC}$ za vsako izmed predstavitev zvočnih posnetkov S_i. Na podlagi teh ocen določimo pragove θ_{S_i} za posamične segmentacije predstavitev S_i po formuli (4.5). Pragove θ_{S_i} in ocene kriterijev BIC, $d^{S_i}_{BIC}$, ki jih računamo iz enačbe (4.3) v osnovnem postopku segmentacije *refBIC*, normaliziramo po obrazcu:

$$norm\_d^{S_i}_{BIC} = \frac{d^{S_i}_{BIC} - minBIC^{S_i}_{DISTBIC}}{maxBIC^{S_i}_{DISTBIC} - minBIC^{S_i}_{DISTBIC}}. \tag{4.6}$$

Podobno dobimo tudi normalizirane pragove $norm\_\theta_{S_i}$, le da v zgornjem obrazcu zamenjamo $d^{S_i}_{BIC}$ s θ_{S_i}.

V drugi fazi postopka izvajamo segmentacijo na podlagi referenčnega postopka *refBIC*, le da se tu zaradi združevanja ocen spremenita korak 2 in 3 osnovnega postopka opisanega v razdelku 4.3.1. Združevanje ocen različnih predstavitev izvedemo kot uteženo vsoto normaliziranih vrednosti kriterijev BIC, $norm\_d^{S_i}_{BIC}$, posamičnih predstavitev S_i:

$$fus\_d_{BIC} = \sum_{i=1}^{S} fw_i \cdot norm\_d^{S_i}_{BIC}. \tag{4.7}$$

fw_i predstavljajo uteži združevanja posamičnih predstavitev S_i in ustrezajo pogoju $\sum_{i=1}^{S} fw_i = 1$, kjer je S število vseh predstavitev. Tako se v koraku 2 metode *refBIC* izvaja iskanje točke t, kjer vrednosti fus_d_{BIC} dosežejo maksimum na intervalu $[a, b]$. Med kandidati t za mejo med segmenti sprejmemo v koraku 3 osnovne metode tistega,

za katerega je $fus\_d_{BIC} > fus\_\theta$, kjer je

$$fus\_\theta = \sum_{i=1}^{S} fw_i \cdot \theta_{S_i}. \tag{4.8}$$

Če velja $fus\_d_{BIC} < fus\_\theta$, kandidata v točki t zavrnemo in sledimo korakom osnovnega postopka naprej.

Za normalizacijo ocen pri združevanju smo izbrali postopek min–max normalizacije, ki pa ne velja za najbolj učinkovitega med postopki normalizacije ocen predvsem zato, ker je zelo občutljiv na napake meritev (*ang. outlier's scores*), [Jain-05]. Za napake veljajo namreč tiste ocene meritev, ki se porazdeljujejo drugače kot večina vrednosti ocen in so ponavadi na robu intervala zaupanja v ocene. Takšne so v večini primerov maksimalne in minimalne vrednosti ocen, zato jih v večini postopkov normalizacije hočemo odpraviti. V našem primeru pa nam ravno maksimalne in minimalne ocene meritev predstavljajo osnovo za izračun pragov odločitve, zato smo temu dejstvu prilagodili tudi normalizacijo ocen.

Preizkušali smo fuzijo dveh predstavitev zvočnih signalov: osnovnih značilk MFCC in njihovih odvodov ΔMFCC, [Young-04]. V tem primeru je bil torej $S = 2$, imeli smo dve predstavitvi S_1 in S_2 ter eno utež združevanja fw_1, saj je bila druga določena s prvo, $fw_2 = 1 - fw_1$. Značilke MFCC in ΔMFCC smo izbrali za segmentacijo iz več razlogov. Prvi je bil v tem, da se značilke MFCC skupaj s kombinacijo ΔMFCC skoraj vedno uporabljajo v postopkih segmentacije. V naših preizkusih se je izkazalo, da smo tudi samo z uporabo značilk ΔMFCC dobili zelo primerljive rezultate segmentacije kot samo z uporabo značilk MFCC. S primerjalnimi testi smo tudi ugotovili, da s statičnimi značilkami MFCC bolje detektiramo meje med segmenti, kjer se spreminjata bodisi akustično ozadje v zvočnem posnetku ali pa se zamenja govorec, medtem ko samo z ΔMFCC bolje detektiramo zamenjave govorcev ob nespremenjenih akustičnih pogojih. Zato je bila za SG (SAG) segmentacijo smiselna uporaba kombinacije obeh tipov značilk. Običajno se pri standardnih postopkih SG (SAG) segmentacije ravno tako izbere 12 MFCC značilk skupaj z logaritmom kratkočasovne energije in njihove odvode [Young-04]. Tako dobimo 26–dimenzionalne vektorje akustičnih značilk, ki jih uporabljamo za ocenjevanje modelov kriterija BIC iz enačbe (4.4). V primeru normalnih porazdelitev moramo tako oceniti povprečni vektor $\mathbf{m}$ in (polno) kovariančno matriko $\mathbf{\Sigma}$, kar v primeru 26–dimenzionalnih vektorjev pomeni ocenjevanje 351 parametrov. To pomeni, da imamo v primeru kratkih segmentov premalo vzorcev (vektorjev značilk) za dobre ocene parametrov modelov kriterija BIC in posledično slabšo segmentacijo krajših odsekov. To pomanjkljivost odpravlja segmentacija s fuzijo. Ker predstavimo značilke MFCC in ΔMFCC kot dva toka podatkov S_1 in S_2, na vsakem odseku tako ocenjujemo dva povprečna vektorja značilk s še enkrat manjšo dimenzijo in prav tako dve kovariančni matriki, kar v primeru 26–dimenzionalnih vektorjev predstavlja ocenjevanje 182 parametrov. Tako znatno znižamo dimenzijo modelov kriterija BIC in lahko boljše ocenjujemo modele tudi v primeru krajših segmentov.

4.5 Preizkusi postopkov segmentacije

V naslednjih razdelkih so opisani preizkusi SAG segmentacije, ki smo jih izvedli v okviru doktorskega dela. Preizkušali smo vse štiri predstavljene postopke segmentacije: osnovni postopek (*refBIC*), opisan v razdelku 4.3.1, postopek *DISTBIC*, opisan v razdelku 4.3.2, postopek segmentacije z relativnim pragom (*relpragBIC*), opisan v razdelku 4.4.1, in postopek segmentacije s fuzijo (*fuzBIC*), opisan v razdelku 4.4.2.

Vrednotenje postopkov je potekalo na dveh zbirkah zvočnih posnetkov informativnih oddaj, SiBN in COST278, predstavljenih v poglavju 2. Medtem ko je zbirka SiBN razmeroma enotna, v njej so namreč zbrani posnetki TV dnevnikov informativnih oddaj v slovenskem jeziku ene TV postaje, pa so v večjezični zbirki COST278 zbrani posnetki različnih informativnih oddaj iz različnih TV postaj v različnih jezikih. Tako smo v primeru zbirke SiBN primerjali učinkovitost postopkov v sorazmerno stabilnih zvočnih pogojih, z zbirko COST278 pa smo želeli ocenjevati robustnost postopkov segmentacije. V ta namen smo del zbirke SiBN uporabili kot razvojno zbirko za ocenjevanje vseh odprtih parametrov postopkov in za natančnejše analize postopkov segmentacije. Optimalno ocenjene parametre iz razvojne zbirke smo potem uporabili v vseh preizkusih segmentacij na zbirkah SiBN in COST278. V primeru SiBN smo tako izvajali postopke segmentacije na podlagi optimalnih izbir parametrov, medtem ko so bili parametri v primeru postopkov segmentacije na zbirki COST278 zaradi tega neoptimalno določeni.

V nadaljevanju bomo naprej opisali mere za vrednotenje postopkov segmentacije, nadaljevali pa z opisom izvedbe preizkusov segmentacije na razvojni in na obeh testnih zbirkah.

4.5.1 Vrednotenje postopkov segmentacije

V postopkih SAG segmentacije merimo običajno dva tipa napak. Napake prve vrste predstavljajo meje med dejanskimi segmenti, ki jih s postopkom segmentacije ne uspemo detektirati. Napake druge vrste pa predstavljajo meje, ki jih s postopkom segmentacije postavimo, vendar jih v dejanskih posnetkih ni.

Napake prve vrste relativno izrazimo s *priklicem (ang. recall, RCL)*, napake druge vrste pa z *natančnostjo (ang. precision, PRC)*, [Kemp-00]:

$$RCL = \frac{\text{št. pravilno določenih mej}}{\text{št. napovedanih mej}} \tag{4.9}$$

$$PRC = \frac{\text{št. pravilno določenih mej}}{\text{št. dejanskih mej}}. \tag{4.10}$$

Z mero *priklica (RCL)* ocenimo, koliko mej bi potrebovali, da bi pravilno napovedali dejanske meje. Mera *natančnosti (PRC)* pa nam podaja delež dejanskih mej, ki bi jih s segmentacijo pravilno določili. Običajno je v postopkih segmentacije priklic večji od natančnosti. Zaradi lažjega vrednotenja postopkov segmentacije pa se uporablja skupna mera napake, *mera F (ang. F–measure)*, ki je sestavljena iz obeh napak:

$$F = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL}. \tag{4.11}$$

Vrednosti *mere F* so na intervalu med 0.0 in 1.0, večje vrednosti pomenijo večje ujemanje mej iz postopka segmentacije z dejanskimi mejami med segmenti.

Pri obeh vrstah napak imamo opravka s pravilno določenimi mejami. Če označimo mejo, ki smo jo dobili s postopkom segmentacije, s t' in dejansko mejo med segmenti s t, potem pravimo, da se meji t in t' ujemata, oziroma da je meja t' pravilno določena, če ustreza pogoju $t-\Delta t \leq t' \leq t+\Delta t$. Pri tem moramo poudariti, da štejemo za pravilno ujemanje samo en par mej (t', t), ostale meje, ki lahko tudi ustrezajo prejšnjemu pogoju pri dani meji t, pa štejemo kot napake. Interval ujemanja določen z Δt je podan vnaprej. V naših preizkusih je bil $\Delta t = 1.0\,\mathrm{s}$.

Običajno je v postopkih vrednotenja segmentacije poleg Δt vhodni parameter tudi predpisana dolžina najkrajšega ne–govornega odseka. Razlog je predvsem v tem, da se moramo v primeru ročno določene segmentacije odločiti, kakšen najmanjši odsek ne–govora bomo še označevali, sicer bi lahko npr. še tako kratko pavzo pri zamenjavi dveh govorcev označili kot ne–govorni segment in tako postavili dve meji pri samo eni zamenjavi govorcev, kar bi povzročilo nepravilno vrednotenje segmentacije. Seveda moramo potem predpisano dolžino najkrajšega ne–govornega odseka upoštevati tudi v postopkih samodejnih segmentacij. V našem primeru je bil najkrajši dovoljeni ne–govorni odsek postavljen na 1.5 s, govorni odseki pa so bili lahko poljubno dolgi.

Vrednotenje postopkov segmentacije smo izvajali z evaluacijskim orodjem, ki je bilo razvito v okviru projekta COST278 za namene evaluacije postopkov samodejne segmentacije in rojenja informativnih oddaj zbirke COST278 [Žibert-05].

4.5.2 Izvedba preizkusov segmentacije

V naših preizkusih smo se omejili predvsem na SAG segmentacijo. Preizkušali smo štiri postopke segmentacije: referenčna postopka *refBIC* in *DISTBIC* ter predlagana postopka *relpragBIC* in *fuzBIC*. V nadaljevanju bomo opisali, kako so bili preizkusi zastavljeni, kako smo določali odprte parametre postopkov in kako smo primerjali postopke med seboj.

V okviru doktorskega dela smo v postopkih segmentacije uporabljali samo akustične predstavitve signalov zvočnih posnetkov, čeprav smo raziskovali tudi drugačne predstavitve govornih signalov. Tako smo npr. razvili značilke na podlagi prozodičnih parametrov govora po zgledu GNG segmentacije, predstavljene v poglavju 3, vendar v primeru SAG segmentacije nismo dosegli primerljivih rezultatov kot ob uporabi samo akustičnih značilk. Preizkuse SAG segmentacije smo tako izvajali na podlagi značilk koeficientov melodičnega kepstra (MFCC) in kratkočasovne energije signalov v kombinaciji z njihovimi odvodi (ΔMFCC), ki se tudi običajno uporabljajo v različnih sistemih za obdelavo govornih signalov. V vseh postopkih segmentacije smo tako uporabljali 12 osnovnih MFCC značilk in logaritem energije (vse skupaj označujemo z MFCC) v kombinaciji z njihovimi odvodi, ki so ponavadi predstavljeni s koeficienti regresijskih premic v okolici osnovnih značilk. Vektorji značilk so bili izračunani na vsakih 10 ms. Tako smo v primeru preizkusov segmentacije s samo osnovnimi značilkami izvajali segmentacije na podlagi 13–dimenzionalnih vektorjev značilk, skupaj z

odvodi pa na podlagi 26–dimenzionalnih vektorjev značilk. Pridobivanje značilk smo izvajali s postopki iz zbirke orodij HTK Toolkit [Young-04], ki je namenjena izgradnji razpoznavalnikov govora.

Izbira dimenzije vektorja značilk je bila bistvenega pomena za dobro ocenjevanje vrednosti kriterija BIC, ki je bil uporabljen v vseh naših preizkusih. V našem primeru smo vse modele uporabljene pri kriteriju BIC opisovali z normalnimi porazdelitvami, torej z eno Gaussovo porazdelitvijo na model, ocenjeno s povprečnim vektorjem in polno kovariančno matriko. V tem primeru smo tako uporabili ocene za kriterijsko funkcijo d_{BIC} izračunane po formuli (4.4). Nekateri avtorji [Ajmera-03, Mori-01, Moraru-03a] so v svojih postopkih segmentacije uporabljali samo diagonalne kovariančne matrike v oceni Gaussovih porazdelitev, vendar smo v naših preizkusih s takim načinom modeliranja dosegli slabše rezultate. V primeru polnih kovariančnih matrik in večje dimenzije vektorjev značilk pa smo morali skrbno izbirati dolžine najkrajših odsekov analize, saj bi v nasprotnem primeru slabo ocenjevali parametre normalnih porazdelitev, posledično dobili slabe ocene kriterija BIC in s tem povzročili številne napake v segmentaciji. Zato je bilo potrebno vnaprej predpisati najkrajši možen segment, ki ga s postopkom segmentacije še lahko zaznamo. V naših eksperimentih smo pri vseh postopkih izbrali najkrajše okno analize dolžine 1.0 s (kolikor je znašal tudi interval ujemanja mej pri vrednotenju postopkov segmentacije).

Vse druge odprte parametre postopkov smo določali za vsak postopek posebej glede na optimalne rezultate segmentacije na razvojni zbirki. V nadaljevanju bomo predstavili, s kakšnimi značilkami smo eksperimentirali in kako smo določali parametre za vsak postopek:

refBIC: Izvajali smo segmentacijo z dvema predstavitvama: osnovnimi značilkami MFCC in z MFCC+ΔMFCC.

V postopku je bilo potrebno določiti še `MIN_ODSEK`, `DODANI_VZORCI` in `MAX_ODSEK`. V primeru obeh predstavitev smo postavili vrednosti na: `MIN_ODSEK` = 2.0 s, `MAX_ODSEK` = 10.0 s, `DODANI_VZORCI` = 100. V primeru koraka izračuna vektorjev značilk na vsakih 10 ms pomeni zadnja vrednost 100 vektorjev ali 1.0 s. λ je bila določena na podlagi optimalnih rezultatov segmentacije na razvojni zbirki.

DISTBIC: Izvajali smo segmentacijo z eno predstavitvijo: MFCC+ΔMFCC.

λ v kriteriju d_{BIC} je bil izbran na podlagi optimalne λ iz postopka *refBIC*; dolžino levega (L_odsek) in desnega odseka (R_odsek) v prvi fazi postopka DISTBIC (slika 4.2) smo postavili na 1.2 s.

Pri iskanju lokalnih maksimumov za kandidate za meje nastopata še dva parametra: minimalna razlika v času med dvema lokalnima maksimumoma in prag θ_{DB}, nad katerim sploh iščemo lokalne maksimume. Minimalno razliko smo postavili na 1.0 s (kot je dolžina najkrajšega odseka), prag θ_{DB} pa smo določili na podlagi rezultatov segmentacije na razvojni zbirki.

relpragBIC: Izvajali smo segmentacijo z dvema predstavitvama: osnovnimi značilkami MFCC in samo z ΔMFCC.

Za oceno vrednosti kriterija BIC v obeh primerih predstavitev smo v prvi fazi uporabili uporabili postopek DISTBIC s parametroma $L\_odsek = R\_odsek = 2.0$ s. Za izbiro praga θ iz enačbe (4.5) je bil $\alpha = 0.05$, torej izbrali smo 5. percentil razlike med maksimalno in minimalno vrednostjo ocene iz postopka DISTBIC.

V drugi fazi postopka smo uporabil enake vrednosti postopka *refBIC* kot v gornjem primeru. Tudi tu smo ocenjevali λ iz razvojne zbirke v obeh primerih predstavitev.

fuzBIC: Izvajali smo segmentacijo z združevanjem segmentacije na podlagi MFCC in segmentacije na podlagi ΔMFCC.

Vsi odprti parametri postopka v primeru obeh predstavitev MFCC in ΔMFCC so bili določeni na podlagi optimalnih vrednosti iz postopka *relpragBIC*.

Na razvojni zbirki smo določali samo parameter uteži združevanja fw_1 iz enačbe (4.7).

Opisani parametri postopkov segmentacije so bili izbrani v glavnem na podlagi preizkusov na razvojni zbirki posnetkov. Pri referenčnih postopkih smo večino parametrov povzeli iz sorodnih eksperimentov segmentacij s temi postopki, [Ajmera-04, Delacourt-01]. Sicer pa izbrani parametri postopkov (razen λ) predstavljajo robne pogoje postopkov segmentacije, v glavnem definirajo področja analize odsekov segmentacije in so intuitivno določljivi. Ostale parametre, kot so pragovi odločitve, uteži fuzije ipd., pa smo izbrali glede na optimalne rezultate segmentacije na razvojni zbirki in jih bomo podrobneje analizirali v naslednjem razdelku.

Omeniti moramo še, da smo v primeru postopka *relpragBIC* testirali dve predstavitvi: MFCC in ΔMFCC. To smo storili zaradi dveh razlogov. Prvi je bil ta, da smo želeli v primeru relativno izbranega praga odločitve pokazati neobčutljivost predlagane segmentacije glede na izbiro predstavitve in izbiro λ kriterija BIC. Drugi razlog pa je bil v tem, da smo obe predstavitvi kasneje uporabili tudi v postopku segmentacije s fuzijo *fuzBIC*. S tem smo lahko v naših preizkusih neposredno primerjali segmentacijo *fuzBIC* s segmentacijo *refBIC* v primeru MFCC+ΔMFCC značilk in hkrati analizirali napredek postopka *fuzBIC* v primerjavi s postopkom *relpragBIC* v primeru segmentacij na podlagi obeh predstavitev.

V vseh naših preizkusih smo tako preizkušali 6 postopkov segmentacij: *refBIC* v primeru MFCC značilk (refBIC: MFCC), *refBIC* v primeru MFCC in ΔMFCC značilk (refBIC: MFCC+ΔMFCC), *DISTBIC* v primeru MFCC in ΔMFCC značilk (DISTBIC: MFCC+ΔMFCC), *relpragBIC* v primeru MFCC značilk (relpragBIC: MFCC) in v primeru ΔMFCC značilk (relpragBIC: ΔMFCC) ter postopek *fuzBIC* z MFCC in ΔMFCC značilkami (fuzBIC: MFCC+ΔMFCC). V vseh primerih so bile segmentacije izvedene brez kakršnekoli vnaprej podane informacije o stanju obdelovanih posnetkov, torej tudi brez predhodne razdelitve posnetkov na govorne in ne–govorne dele.

Vsi postopki so bili izvedeni z lastno razvitimi orodji.

4.5.3 Primerjava postopkov segmentacije na razvojni zbirki

Postopke segmentacije na razvojni zbirki smo izvajali zaradi dveh razlogov. Osnovni razlog, ki smo ga že omenili, je bil, da smo na podlagi optimalnih rezultatov segmentacije na razvojni zbirki ocenili ključne parametre vseh postopkov segmentacij. Drugi razlog pa je bil, da smo lahko na podlagi razvojne zbirke neposredno primerjali in ovrednotili vse postopke segmentacije na celotnih območjih njihovega delovanja. S tem smo lahko sklepali na učinkovitost preizkušanih postopkov ob spremenljivih pogojih delovanja.

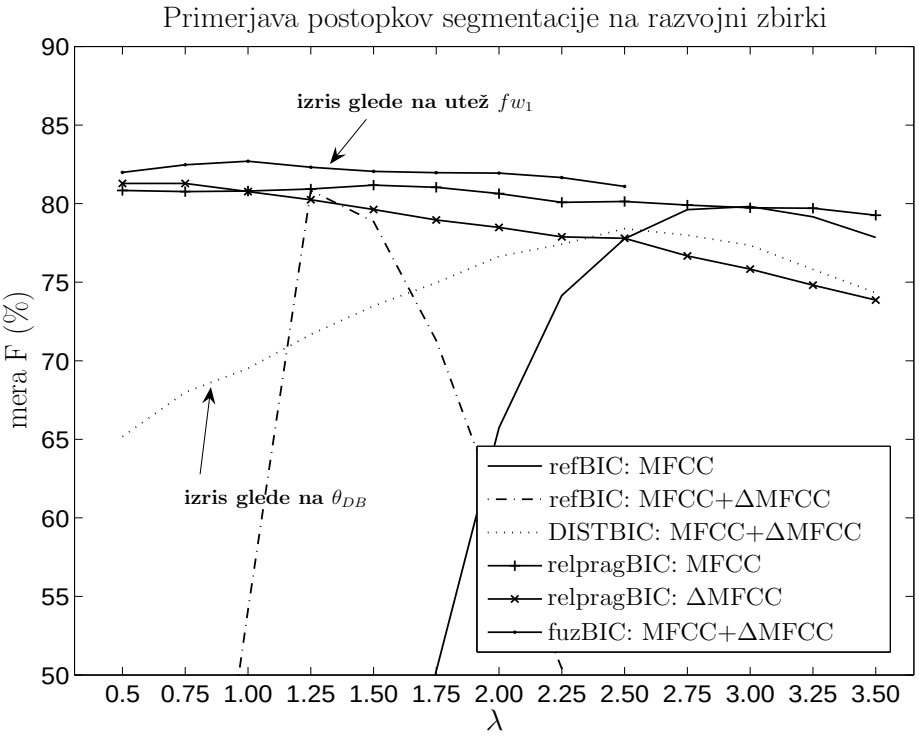
Razvojno zbirko smo sestavili iz posnetkov sedmih informativnih oddaj iz zbirke SiBN v skupnem trajanju okoli 7 ur. Število mej med segmenti v SAG segmentaciji, ki jih je bilo potrebno detektirati, je bilo 1767. Razvojna zbirka je bila tako sestavljena iz razmeroma enovitih posnetkov informativnih oddaj ene TV postaje v slovenskem jeziku. Namenoma je bila zasnovana tako, da bi z njo optimirali delovanje postopkov segmentacije na zbirki SiBN, hkrati pa preverili delovanje postopkov na zbirki COST278 ob neoptimalno izbranih parametrih postopkov. S tem smo hoteli še dodatno preizkusiti delovanje postopkov segmentacije ob optimalnih in neoptimalnih pogojih delovanja. Razdelitev posnetkov iz zbirk SiBN in COST278 med razvojne in testne posnetke je natančnejše opisana v dodatku A disertacije.

Optimizacijo postopkov segmentacije smo izvajali na podlagi najboljših rezultatov segmentacije, merjenih z *mero F* iz formule (4.11).

Določali smo optimalne vrednosti različnih parametrov. V postopkih *refBIC* in *relpragBIC* smo iskali optimalno vrednost λ kriterija BIC. Optimalni parameter λ smo iskali na intervalu od 0.50 do 3.50 s korakom 0.25. Dodatno smo v postopkih *relpragBIC* preverjali tudi izbiro relativnega praga α. Izbirali smo med vrednostmi 0.05, 0.10 in 0.15. Skoraj pri vseh izbranih λ smo ob izbiri $\alpha = 0.05$ dosegli praviloma najboljše rezultate segmentacije, zato smo α v vseh postopkih segmentacije postavili na 0.05. V postopku *DISTBIC* je bila λ izbrana glede na optimalne rezultate segmentacije s postopkom *refBIC: MFCC+ΔMFCC*. Tu smo iskali optimalen prag θ_{DB} odločitve za mejo med segmentoma, uporabljen v drugi fazi postopka *DISTBIC*. Vrednosti θ_{DB} smo izbirali iz intervala $[-300, 0]$ s korakom 25. Pri postopku *fuzBIC* smo iskali optimalne vrednosti uteži združevanja fw_1 in $fw_2 = 1 - fw_1$ za predstavitve MFCC in ΔMFCC. Vsi ostali parametri, λ in α iz obeh predstavitev, so bili določeni glede na najboljše rezultate segmentacij s postopkoma *relpragBIC: MFCC* in *relpragBIC: ΔMFCC*. Iskanje optimalne segmentacije je bilo izvedeno za uteži od 0.1 do 0.9 s korakom iskanja 0.1.

Na sliki 4.5 so grafično prikazani rezultati segmentacije v primeru vseh šestih postopkov, v tabeli 4.1 pa so zbrani najboljši rezultati posameznih segmentacij ob izbiri optimalnih parametrov postopkov.

Na sliki 4.5 so prikazani vsi rezultati segmentacij na podlagi *mere F* v odvisnosti od spreminjanja parametrov posameznih metod. Grafi rezultatov postopkov *refBIC: MFCC*, *refBIC: MFCC+ΔMFCC*, *relpragBIC: MFCC* in *relpragBIC: ΔMFCC* so izrisani na podlagi spreminjanja vrednosti λ in so med seboj poravnani. Grafa rezultatov postopka *DISTBIC: MFCC+ΔMFCC* in *fuzBIC: MFCC+ΔMFCC* nista poravnana z ostalimi rezultati, saj sta izrisana glede na spreminjanje praga odločitve θ_{DB} v prvem



Slika 4.5: Primerjava postopkov segmentacije na razvojni zbirki. Pri postopkih *refBIC* in *relpragBIC* je prikazan graf spreminjanja *mere F* glede spreminjanje λ, pri postopku *DISTBIC* so prikazane spremembe glede na izbiro praga odločitve θ_{DB} za mejo v drugi fazi postopka, v primeru *fuzBIC* pa je podan prikaz odvisnosti *mere F* od uteži fuzije fw_1.

oziroma glede na spreminjanje uteži fw_1 v drugem primeru. Na sliki 4.5 so vsi rezultati zbrani skupaj zaradi lažje primerjave postopkov segmentacije med seboj.

Iz prikaza rezultatov na sliki 4.5 je razvidno, da smo s postopkom *fuzBIC: MFCC+ΔMFCC* na razvojni zbirki dosegli najboljše rezultate segmentacije. Zelo dobre rezultate smo dosegli tudi s postopkom *relpragBIC: MFCC*, vendar so bili nekaj slabši kot v primeru segmentacije s fuzijo. Presenetljivo dobri rezultati so bili doseženi tudi v primeru segmentacije s postopkom *relpragBIC: ΔMFCC*. Tu smo samo z odvodi osnovnih značilk MFCC ob izbiri λ med 0.5 in 1.0 celo presegli rezultate segmentacije s postopkom *relpragBIC* v primeru osnovnih značilk. Z referenčnimi postopki *refBIC: MFCC*, *refBIC: MFCC+ΔMFCC* in *DISTBIC: MFCC+ΔMFCC* smo praviloma dosegli slabše rezultate segmentacije, vendar smo bili v določenih točkah delovanja postopkov blizu najboljšim rezultatom. Iz grafa rezultatov segmentacij pa je predvsem razvidna občutljivost delovanja referenčnih metod na spremembe odprtih parametrov postopkov. Tako lahko v primeru vseh referenčnih postopkov opazimo precej strma naraščanja in padanja rezultatov v bližini optimalnih vrednosti parametrov, še posebej je to razvidno v primeru postopka *refBIC: MFCC+ΔMFCC*. Če slabo določimo parametre takšnih postopkov, lahko dobimo zelo slabe rezultate segmentacij. V primeru predlaganih postopkov *relpragBIC* in *fuzBIC* pa lahko opazimo stabilno delovanje segmentacij ne glede na izbiro parametrov in predstavitev. Zaradi tega lahko pričakujemo sorazmerno stabilno delovanje takšnih postopkov tudi v primeru neznanih pogojev segmentacije in

različnih akustičnih lastnosti zvočnih posnetkov.

Tabela 4.1: Rezultati postopkov SAG segmentacije na razvojni zbirki ob izbiri optimalnih parametrov segmentacij.

postopek segmentacije	*optimalni parametri*	*RCL* (%)	*PRC* (%)	*mera F* (%)
refBIC: MFCC	$\lambda = 3.00$	85.30	75.00	79.82
refBIC: MFCC+ΔMFCC	$\lambda = 1.25$	86.60	75.80	80.84
DISTBIC: MFCC+ΔMFCC	$\lambda = 1.25$, $\theta_{DB} = -200$	80.30	76.60	78.40
relpragBIC: MFCC	$\lambda = 1.50$, $\alpha = 0.05$	86.60	76.40	81.18
relpragBIC: ΔMFCC	$\lambda = 0.75$, $\alpha = 0.05$	84.00	78.70	81.29
fuzBIC: MFCC+ΔMFCC	$\lambda = 0.75$, $\alpha = 0.05$, $fw_1 = 0.3$	85.00	80.50	82.70

V tabeli 4.1 so zbrani samo najboljši rezultati SAG segmentacije ob izbiri optimalnih parametrov segmentacije. Tu lahko še enkrat ugotovimo, da smo na razvojni zbirki s predlaganimi postopki presegli rezultate segmentacije referenčnih metod. Rezultati segmentacije s predlaganimi metodami so višji predvsem zaradi višjih vrednosti mere *natančnosti (PRC)* ob nespremenjenem *priklicu (RCL)*. Za relevantnost vrednotenja postopkov segmentacije je potrebno seveda primerjati postopke glede na nabor značilk, ki smo jih uporabili v segmentaciji. V tem primeru moramo tako primerjati postopek *refBIC: MFCC* s postopkoma *relpragBIC: MFCC* in *relpragBIC: ΔMFCC* ter postopke *refBIC: MFCC+ΔMFCC*, *DISTBIC: MFCC+ΔMFCC* in *fuzBIC: MFCC+ΔMFCC* med seboj. V prvem primeru lahko ugotovimo, da smo v primeru predlaganih postopkov popravili segmentacijo za 1.4% absolutno ob optimalnih izbirah parametrov. V drugem primeru pa smo z uporabo postopka *fuzBIC* pridobili slaba 2% v primerjavi z *refBIC* postopkom in 4.3% v primeru *DISTBIC* postopka.

Optimalne parametre postopkov segmentacij, ki so prikazani v tabeli 4.1, smo uporabili v ostalih preizkusih segmentacije na zbirkah SiBN in COST278 in jih bomo opisali v nadaljevanju.

4.5.4 Primerjava postopkov segmentacije na testnih zbirkah

Vrednotenje postopkov segmentacije v večjem obsegu smo izvajali na dveh testnih zbirkah zvočnih posnetkov informativnih oddaj, zbirki SiBN in COST278, ki sta bili opisani v poglavju 2. Skupno število posnetkov obeh zbirk je bilo 73 v skupnem trajanju okoli 57 ur. Preizkuse segmentacije smo izvajali ločeno na obeh zbirkah, da bi tako primerjali delovanje postopkov segmentacije ob izbiri optimalnih parametrov ob nespremenjenih pogojih zvočnih posnetkov v primeru SiBN zbirke in delovanje postopkov ob izbiri neoptimalnih parametrov v primeru večjezične zbirke COST278.

4.5.4.1 Primerjava postopkov segmentacije na zbirki SiBN

Vrednotenje postopkov segmentacije v primeru SiBN zbirke smo izvajali na tistih posnetkih, ki jih nismo uporabili v razvojni zbirki. Število posnetkov v testnem delu je bilo tako 27 v skupnem trajanju okoli 27 ur. V primeru SAG segmentacije je bilo tako potrebno določiti 6782 mej med segmenti.

Testirali smo vseh šest postopkov segmentacije z izbranimi parametri, ki smo jih določili na podlagi najboljših rezultatov segmentacij na razvojni zbirki.

Tabela 4.2: Rezultati postopkov SAG segmentacije na zbirki SiBN ob izbiri optimalnih parametrov glede na razvojno zbirko.

postopek segmentacije	*RCL* (%)	*PRC* (%)	*mera F* (%)
refBIC: MFCC	88.3	74.3	80.7
refBIC: MFCC+ΔMFCC	86.6	76.6	81.3
DISTBIC: MFCC+ΔMFCC	80.0	79.0	79.5
relpragBIC: MFCC	85.8	77.4	81.4
relpragBIC: ΔMFCC	84.2	80.1	82.1
fuzBIC: MFCC+ΔMFCC	84.4	81.7	83.0

V tabeli 4.2 so zbrani vsi rezultati postopkov segmentacije. Kot lahko ugotovimo, so se razmerja med rezultati vseh segmentacij ohranila in se niso veliko spremenila v primerjavi z rezultati iz razvojne zbirke. To pomeni, da smo parametre postopkov iz razvojne zbirke zelo dobro ocenili, kar je bilo tudi pričakovati. Tudi tu smo najboljše rezultate dosegli v primeru segmentacije s fuzijo (*fuzBIC:MFCC+ΔMFCC*) zaradi dobrega razmerja med *RCL* in *PRC*. Omeniti velja še, da je bil drugi najvišji rezultat segmentacije dosežen s postopkom *relpragBIC* v primeru odvodov značilk ΔMFCC, kar je precej presenetljivo, saj se predstavitve zvočnih posnetkov samo z odvodi osnovnih značilk skoraj nikoli ne uporabljajo v postopkih segmentacije.

Tudi tu lahko ugotovimo, da so skupni rezultati segmentacij z našimi predlaganimi postopki višji od referenčnih metod. Glede na število vseh mej med segmenti, ki smo jih morali odkriti, in glede na izbiro optimalnih parametrov pri vseh postopkih segmentacije, lahko trdimo, da predlagani postopki predstavljajo dejanski napredek v segmentaciji.

4.5.4.2 Primerjava postopkov segmentacije na zbirki COST278

V primeru preizkusov segmentacije na večjezični zbirki COST278 smo hoteli primerjati delovanje postopkov v spremenljivih razmerah zvočnih posnetkov različnih TV oddaj ob neoptimalno izbranih parametrih postopkov segmentacije.

Preizkuse smo izvajali na 56-ih posnetkih informativnih oddaj v skupnem trajanju okoli 30 ur. V primeru SAG segmentacije je bilo potrebno detektirati 8012 mej med segmenti.

Tabela 4.3: Rezultati postopkov SAG segmentacije na zbirki COST278 ob izbiri optimalnih parametrov glede na razvojno zbirko.

postopek segmentacije	*RCL* (%)	*PRC* (%)	*mera F* (%)
refBIC: MFCC	86.0	64.7	73.9
refBIC: MFCC+ΔMFCC	68.7	78.0	73.1
DISTBIC: MFCC+ΔMFCC	70.6	70.3	70.5
relpragBIC: MFCC	84.1	69.8	76.3
relpragBIC: ΔMFCC	78.9	67.7	72.9
fuzBIC: MFCC+ΔMFCC	82.1	72.3	76.9

V tabeli 4.3 so prikazani rezultati vseh šestih segmentacij. Tu so se razmerja med rezultati segmentacije nekoliko spremenila, kar je posledica neoptimalne izbire parametrov. To je lepo razvidno v primeru postopka *refBIC: MFCC+ΔMFCC*, saj so se tu rezultati glede na prejšnje preizkuse zelo poslabšali. Predvsem se je zelo spremenilo razmerje med *RCL* in *PRC*, kar govori v prid dejstvu, da je metoda *refBIC* v primeru večje dimenzije značilk predstavitve posnetkov za segmentacijo zelo občutljiva na izbiro praga odločitve λ. To je lepo razvidno tudi iz slike 4.5, kjer postopek *refBIC: MFCC+ΔMFCC* dobro deluje le na zelo ozkem področju izbire parametra λ. Tako v primeru segmentacije s postopkom *fuzBIC: MFCC+ΔMFCC* dobimo za 3.8 % boljše rezultate absolutno kot s postopkom *refBIC: MFCC+ΔMFCC* in 6.4 % boljše od postopka *DISTBIC: MFCC+ΔMFCC*. Tudi razlika v rezultatih segmentacije s postopkom *refBIC: MFCC* in *relpragBIC: MFCC* se je povečala, medtem ko so se rezultati segmentacije s postopkom *relpragBIC: ΔMFCC* poslabšali. Rezultati segmentacije s postopkom *DISTBIC: MFCC+ΔMFCC* so bili v vseh preizkusih nekoliko slabši od referenčnega postopka *refBIC*. To je predvsem zaradi slabšega določanja lokalnih maksimumov, ki smo se jim v primeru predlaganih postopkov *relpragBIC* in *fuzBIC* uspešno izognili.

Glede na rezultate segmentacije v primeru razvojne in testnih zbirk lahko ugotovimo, da so predlagani postopki delovali v povprečju boljše od referenčnih metod. Predvsem pa lahko zaključimo, da so bili bolj neodvisni od izbire začetnih odprtih parametrov postopkov, zlasti od izbire pragov odločitve za meje med segmenti.

4.6 Zaključek

V tem poglavju smo se ukvarjali s segmentacijo zvočnih posnetkov informativnih oddaj glede na zamenjave govorcev in/ali spremembe v akustičnem ozadju. Razdelitev večjih posnetkov na manjše odseke - segmente - je ključnega pomena za nadaljnjo obdelavo posnetkov v sistemih za samodejno razpoznavanje govora in v sistemih, ki temeljijo na razpoznavanju govorcev. Postopki segmentacije morajo biti zato učinkoviti, da bistveno ne poslabšajo delovanje ostalih postopkov v sistemih za obdelavo govora, hkrati pa morajo biti dovolj neobčutljivi na spremembe zvočnih razmer v zajetih posnetkih, saj običajno delujejo v prvih fazah obdelave podatkov v omenjenih sistemih.

V okviru doktorskega dela smo poskušali izboljšati postopke segmentacije zvočnih posnetkov v primeru zamenjave govorcev in sprememb v akustičnem ozadju. Pri tem smo se omejili na postopke segmentacije s kriterijem BIC, ki trenutno velja za najboljšega med kriteriji za določitev mej med segmenti. Preučili in ovrednotili smo dva referenčna postopka segmentacije na podlagi kriterija BIC ter predlagali dva nova postopka: postopek segmentacije na podlagi relativno določenega praga odločitve za mejo in postopek segmentacije z združevanjem različnih predstavitev segmentacije. V obeh predlaganih postopkih smo združili posamezne faze obeh predstavljenih referenčnih metod. Glavna ideja je bila v tem, da smo iz referenčnega postopka DISTBIC ocenili vrednosti kriterija BIC za drugi referenčni postopek. Na ta način smo pridobili informacijo o porazdelitvi ocen kriterija BIC, ki smo jo v prvem primeru uporabili za sprotno določitev pragov odločitve za meje v posameznem posnetku, v drugem pa za normalizacijo vrednosti ocen, kar nam je omogočilo združevanje različnih postopkov segmentacije.

Primerjava referenčnih in predlaganih postopkov segmentacije je bila izvedena na razvojni zbirki in testnih zbirkah SiBN in COST278. V preizkusih se je izkazalo, da so predlagane metode bolj učinkovite od referenčnih, predvsem pa bolj neobčutljive na spremembe v posnetkih in na spremenjene pogoje delovanja segmentacij. To je bil tudi naš cilj pri načrtovanju predlaganih postopkov. Z uvedbo relativno določenega praga odločitve za meje med segmenti smo prevedli izbiro absolutnih vrednosti pragov v relativne izbire, ki se spreminjajo glede na posnetke, ki jih obdelujemo. S tem smo pridobili sorazmerno neobčutljive pragove odločitve za meje med segmenti in tako na drugačen način rešili problem preobčutljivosti postopkov segmentacije na izbiro odprtih parametrov. Stranski učinek ocenjevanja vrednosti kriterija BIC iz postopka DISTBIC pa je bila možna normalizacija pričakovanih ocen kriterija. Tako smo lahko z normalizacijo ocen in relativno ocenjenimi pragovi odločitve izvajali segmentacijo na podlagi združevanja uteženih ocen odločitve kriterija BIC iz več predstavitev zvočnih posnetkov. Na ta način smo v vseh preizkusih segmentacije dosegli najboljše rezultate.

5 Razvrščanje segmentov po govorcih s postopki rojenja

V tem poglavju se bomo posvetili razvrščanju segmentov zvočnih posnetkov informativnih oddaj glede na govorce. Naloga razvrščanja segmentov po govorcih je, da razdelimo posamezno informativno oddajo na odseke, ki pripadajo samo enemu govorcu in te odseke povežemo skupaj v razred tega govorca. Prvi del problema smo reševali v prejšnjem poglavju s postopki samodejne segmentacije po govorcih. V tem poglavju pa se bomo posvetili postopkom rojenja tako pridobljenih segmentov z namenom razvrščanja segmentov po govorcih v danih posnetkih.

V uvodu bomo natančneje formulirali problem razvrščanja segmentov po govorcih ter v nadaljevanju opisali dva postopka rojenja, ki se najbolj pogosto uporabljata pri razvrščanju segmentov. Glavni problemi, ki jih rešujemo pri razvrščanju segmentov po govorcih s postopki rojenja, so: izbira kriterija združevanja segmentov, predstavitve segmentov in kriteriji za zaustavitev postopkov rojenja.

V našem raziskovalnem delu smo se posvetili predvsem predstavitvam zvočnih posnetkov za združevanje segmentov glede na govorce in kriterijem zaustavitve rojenja. Najboljše obstoječe postopke in predlagane postopke rojenja smo primerjali na obsežnih zbirkah zvočnih posnetkov v primeru ročno označenih segmentov in v primeru samodejno pridobljenih segmentov posnetkov informativnih oddaj.

5.1 Uvod

Pri rojenju vzorcev gre ponavadi za združevanje ali razdruževanje množic vzorcev na podlagi podobnosti ali različnosti med njimi. Pri tem običajno izvajamo postopke združevanja (razdruževanja), dokler se dá oziroma dokler ni izpolnjen nek vnaprej predpisan pogoj zaustavitve postopka rojenja. V splošnem postane problem rojenja dobro definiran šele, ko predpišemo, kakšne roje pričakujemo oziroma kakšne razrede vzorcev naj predstavljajo roji. Lastnosti, ki jih želimo združevati (razdruževati), tako opišemo s predstavitvami vzorcev ali množic vzorcev, ki jih rojimo. Mere podobnosti (različnosti) nam služijo za kriterije združevanja (razdruževanja), kriterije zaustavitve postopkov rojenja pa določimo tako, da bi bila napaka med roji in pričakovanimi razredi najmanjša.

V procesu segmentacije in razvrščanja segmentov po govorcih *(ang. speaker diarisation)* želimo pridobiti in povezati med seboj tiste dele zvočnih posnetkov, ki pripadajo istim govorcem. Za to je potrebno najprej razdeliti zvočne posnetke na ustrezne segmente, ki jih nato lahko povezujemo s postopki rojenja. To pomeni, da je problem segmentacije in razvrščanja sestavljen iz več nalog. V prvi fazi moramo ustrezno razdeliti zvočni posnetek, najprej na govorne in ne–govorne dele, nato pa govorne dele tako, da bodo primerni za povezovanje v drugi fazi. Šele nato s postopki rojenja takšnih odsekov pridobimo informacijo, kje v posnetku se nahaja kakšen govorec, kar je tudi cilj razvrščanja segmentov zvočnih posnetkov. S postopki prve faze oziroma razdelitvijo posnetkov na segmente smo se ukvarjali v poglavjih 3 in 4. V tem poglavju pa se bomo ukvarjali s postopki rojenja segmentov po govorcih *(ang. speaker clustering)*.

Cilj rojenja segmentov po govorcih je, da bi v idealnem primeru enega govorca opisali z enim samim rojem segmentov. To je zaželeno v številnih sistemih govornih tehnologij, ki temeljijo na strukturiranju zvočnih posnetkov po govorcih, kot so npr. sistemi za indeksacijo zvočnih posnetkov, sistemi za iskanje in sledenje govorcev v zvočnih posnetkih, sistemi za razpoznavanje govorcev ipd. Pri drugi skupini govornih sistemov, ki temeljijo na razpoznavanju govora, pa je bolj smiselno združevati govorce v roje na podlagi skupnih akustičnih in prozodičnih lastnosti (kot so tip posnetka, akustično ozadje govora, način govora, hitrost govorjenja, spol, ...). To pa predvsem zaradi tega, ker v takšnih sistemih poteka prilagajanje modelov razpoznavanja na govorčeve lastnosti. V tem primeru lahko uporabimo enake postopke rojenja, le kriterij zaustavitve je potrebno prilagoditi. Tako so številne študije pokazale [Chen-98, Johnson-98, Meinedo-03b, Zhang-02, Beyerlein-02, Woodland-02, Chen-02, Gauvain-03, Viswanathan-00, Makhoul-00] smiselnost uporabe rojenja po govorcih v obeh primerih, v primeru prilagajanja glede na govorce v sistemih za razpoznavanje ter v sistemih indeksacije in iskanja govorcev v multimedijskih zbirkah.

V okviru doktorskega dela smo se tako osredotočili na rojenje segmentov po govorcih (rojenje SG), saj nam je bil cilj strukturirati zvočne posnetke informativnih oddaj na odseke, ki pripadajo istemu govorcu. Osnovna vprašanja, ki smo jih tu reševali, vključujejo iskanje primernih predstavitev segmentov za rojenje po govorcih, izbiro mere podobnosti (različnosti) med segmenti in kriterija zaustavitve rojenja. Osredotočili smo se predvsem na predstavitve segmentov in kriterije zaustavitve postopkov rojenja, medtem ko smo za mero podobnosti med segmenti uporabili kriterij BIC. Ker nam

je bil cilj rojenja, da bi pridobili takšne roje segmentov, ki bi pripadali samo enemu govorcu, smo se pri predstavitvah segmentov in kriterijih združevanja zgledovali po postopkih razpoznavanja govorcev. Tako smo poleg osnovnega postopka rojenja SG, ki temelji na kriteriju BIC, preizkusili in dodatno izboljšali alternativen postopek, ki temelji na metodah razpoznavanja govorcev. Pri predstavitvah segmentov smo poleg akustične informacije vpeljali še prozodično in tako razvili nov postopek rojenja, ki združuje obe predstavitvi. Vse postopke smo preizkušali v dveh primerih: v primeru ročno označenih segmentov, se pravi, v primeru idealne segmentacije, in v primeru samodejno pridobljenih segmentov, torej v primeru samodejne segmentacije.

Ukvarjali smo se tudi s kriteriji zaustavitve postopkov rojenja. Tu smo preizkusili številne že uveljavljene kriterije in predlagali dva nova. Prvi je temeljil na meri podobnosti, ki smo jo uporabljali za združevanje segmentov, drugi pa na meri napake, s katero smo ocenjevali uspešnost delovanja postopkov rojenja. S tem smo hoteli, na eni strani na podlagi postopka rojenja, na drugi strani pa na podlagi mere napake rojenja, oceniti število pričakovanih govorcev oziroma končno število rojev v danem posnetku.

Hkrati smo z razvojem postopkov postavili tudi ogrodje za vrednotenje postopkov razvrščanja segmentov po govorcih. Poleg že omenjene primerjave postopkov ob idealni in samodejni segmentaciji posnetkov smo predlagali tudi drugačno metodo vrednotenja postopkov, kjer smo lahko primerjali postopke ne glede na izbiro kriterija za zaustavitev rojenja.

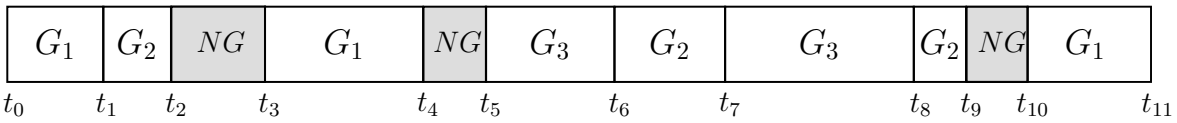
V nadaljevanju bomo najprej bolj natančno definirali problem razvrščanja segmentov po govorcih in opisali dva referenčna postopka rojenja SG. V razdelku 5.4 bomo predstavili predlagani postopek rojenja na podlagi akustične in prozodične informacije, ki temelji na referenčnem postopku ob uteženem združevanju obeh predstavitev. V razdelku 5.5 je podana primerjava vseh postopkov na zbirki SiBN v primeru idealne in samodejne segmentacije posnetkov in na zbirki COST278 samo v primeru samodejne segmentacije. Primerjava postopkov je bila izvedena s predlaganim postopkom vrednotenja brez upoštevanja kriterijev zaustavitve rojenja, s katerimi se ukvarjamo v razdelku 5.6. Ob testiranju različnih kriterijev zaustavitve so tako v zaključnem delu podani končni rezultati razvrščanja segmentov po govorcih v primeru idealne in v primeru samodejne segmentacije na zbirkah SiBN in COST278.

5.2 Formulacija problema

Kot smo že povedali, združuje problem razvrščanja segmentov po govorcih več nalog. Zvočni posnetek je potrebno najprej razdeliti na govorne in ne–govorne dele, nato je potrebno govorne dele ustrezno predstaviti z manjšimi odseki, ki predstavljajo osnovne enote združevanja ali razdruževanja v postopkih rojenja.

Na sliki 5.1 je prikazan končni rezultat razvrščanja segmentov po govorcih. Naloga razvrščanja je torej, da zvočni posnetek razdelimo na segmente, ki pripadajo posameznemu govorcu, jih opremimo z informacijo o začetku in koncu trajanja segmenta in vsak segment ustrezno označimo. Oznaka NG združuje vse segmente, ki predstavljajo ne–govorne dele v posnetku. Oznake G_i pa pripadajo posameznim govorcem v

posnetku. Segmenti, ki pripadajo istemu govorcu, imajo isto oznako.



Slika 5.1: Končni rezultat razvrščanja segmentov po govorcih. Vsak segment je opremljen z informacijo o začetku in koncu segmenta ter z oznako, kateremu govorcu pripada.

Postopek razvrščanja je tako običajno sestavljen iz treh faz. V prvi fazi določimo govorne in ne–govorne dele, v drugi izvedemo segmentacijo po govorcih na govornih delih, v tretji pa izvedemo povezovanje segmentov glede na govorce s postopki rojenja SG. V našem raziskovalnem delu smo sledili temu zaporedju in smo postopke v vseh fazah razvrščanja izvajali med seboj neodvisno. Tako smo sledili standardnemu pristopu reševanja tega problema, čeprav obstajajo tudi druge možne poti predvsem v smeri združevanja segmentacije in rojenja iz druge in tretje faze postopka. V tem primeru osnovne odseke združevanja ne predstavljajo segmenti po govorcih, ampak običajno krajši odseki, ki jih ustrezno spreminjamo (povečujemo ali manjšamo), da dosežemo najboljše pogoje za združevanje. Takšna sta npr. postopka, opisana v [Ajmera-04] in [Zdansky-05]. Prednost takšnih pristopov je predvsem v tem, da se napake segmentacije ne prenašajo naprej v rojenje, slabost pa je v tem, da moramo poleg povezovanja segmentov še določati njihove meje, kar pomeni večjo računsko kompleksnost problema in s tem povečan čas delovanja razvrščanja.

Ker smo sledili standardnemu pristopu razvrščanja, se bomo v tem poglavju ukvarjali z zadnjo fazo, torej s postopki rojenja SG. Da bi bolje ovrednotili postopke rojenja, ki smo jih razvili, smo jih preizkušali v dveh nalogah. V prvi smo predpostavili idealne pogoje segmentacije. To pomeni, da smo izvajali rojenje SG na ročno označenih segmentih zvočnih posnetkov, se pravi, da smo pustili ročno označene ne–govorne dele, v govornih delih pa smo ročno postavili meje med govorci. V tem primeru bomo govorili o idealni segmentaciji. V drugem primeru pa smo izvajali rojenje na samodejno pridobljenih segmentih. Za segmentacijo smo tu uporabili najboljše metode GNG segmentacije, predstavljene v poglavju 3 in SAG segmentacije, opisane v poglavju 4. Pri tem moramo poudariti, da smo v našem primeru najprej izvajali SAG segmentacijo in nato GNG klasifikacijo. Na tako pridobljenih segmentih smo potem izvajali postopke rojenja SG, zato bomo v tem primeru govorili o rojenju SG na podlagi samodejne segmentacije.

Omeniti moramo še sorodnost problema rojenja segmentov po govorcih in segmentacije po govorcih. Medtem ko v prvem primeru ugotavljamo, ali je primerno združiti ali razdružiti dva poljubna segmenta v posnetku, se v drugem sprašujemo, ali je možno postaviti mejo med dvema sosednima odsekoma v posnetku, torej ali je bolje pustiti združena dva sosedna segmenta ali ju je bolje razdružiti - postaviti mejo. Zato se lahko v obeh primerih uporabljajo sorodne metode pri reševanju obeh problemov tako pri kriterijih združevanja, predstavitvah segmentov in kriterijih zaustavitve rojenja oziroma praga odločitve za mejo.

5.3 Referenčni postopki rojenja segmentov

V tem razdelku bosta predstavljena dva postopka rojenja SG, ki se uporabljata pri razvrščanju segmentov zvočnih posnetkov. Prvi je osnovni postopek združevanja segmentov na podlagi kriterija BIC in smo ga uporabljali kot referenčni postopek rojenja SG. Drugi postopek rojenja je izpeljan iz metod, ki se uporabljajo pri razpoznavanju govorcev in predstavlja alternativo osnovnemu pristopu. Oba postopka temeljita na hierarhičnih postopkih iskanja rojev [Pavešić-00, poglavje 5.5.3, str. 242–249].

5.3.1 Osnovni postopek rojenja z združevanjem segmentov

Osnovni postopek rojenja SG z združevanjem segmentov temelji na hierarhičnem postopku iskanja rojev. Postopek združevanja segmentov poteka od spodaj navzgor in ga lahko zapišemo v naslednjih korakih:

- 1. Začetek:
  - 1.1. Vsak segment C_i, $i = 1, \ldots, N$ predstavlja en roj.
    Množica rojev je: $\mathcal{S}_0 = \{C_i \,|\, i = 1, \ldots, N\}$.
  - 1.2. $t = 0$.
- 2. Ponavljaj:
  - 2.1. $t = t + 1$;
  - 2.2. Med vsemi pari rojev (C_r, C_s) v $\mathcal{S}_{t-1}$ najdemo tak par (C_i, C_j), za katerega velja:

$$g(C_i, C_j) = \begin{cases} \min_{r,s} g(C_r, C_s), & \text{če je } g \text{ mera različnosti} \\ \max_{r,s} g(C_r, C_s), & \text{če je } g \text{ mera podobnosti} \end{cases} \tag{5.1}$$

  - 2.3. Nov roj je $C_q = C_i \cup C_j$. Popravi množico rojev
    $\mathcal{S}_t = (\mathcal{S}_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$.
- Dokler ni izpolnjen *kriterij zaustavitve* rojenja.

Shema 5.2: Splošen postopek hierarhičnega rojenja od spodaj navzgor, ki smo ga uporabljali pri rojenju SG.

Postopek rojenja SG iz sheme 5.2 je prirejen po splošni shemi hierarhičnega postopka rojenja z združevanjem *(ang. agglomerative clustering)*, [Theodoridis-03, str. 450-451]. V našem primeru osnovne vzorce rojenja predstavljajo segmenti, ki jih združujemo po principu od spodaj navzgor *(ang. bottom–up clustering)*, se pravi, da v prvem koraku rojenja vsak segment predstavlja en roj C_i, v vsakem naslednjem koraku pa združimo dva roja (C_i, C_j), ki sta si med seboj najbolj podobna ali različna glede na mero združevanja g. Postopek se konča, ko so izpolnjeni pogoji *kriterija zaustavitve* rojenja.

V vseh postopkih, ki bodo opisani v tem poglavju, smo uporabili to shemo rojenja, razlike so bile le v predstavitvah segmentov oziroma rojev C_i, v kriteriju združevanja g in kriterijih zaustavitve rojenja. V prvem delu, kjer bomo primerjali postopke rojenja SG na podlagi različnih predstavitev segmentov in različnih kriterijev združevanja, smo izvajali rojenje do konca, torej brez uporabe kriterija zaustavitve. V drugem delu pa se bomo posvetili kriterijem zaustavitve rojenja.

V osnovnem postopku rojenja SG smo za mero združevanja uporabljali kriterij BIC, ki smo ga opisali že v prejšnjem poglavju v razdelku 4.2.1. V tem primeru smo roje C_i predstavili z modeli normalnih porazdelitev, torej s povprečnim vektorjem $\mathbf{m}_i$ in (polno) kovariančno matriko $\mathbf{\Sigma}_i$. Tako za mero g računamo d_{BIC} iz enačbe (4.4):

$$bic(C_i, C_j) = \frac{N_{ij}}{2} \log |\mathbf{\Sigma}_{ij}| - \frac{N_i}{2} \log |\mathbf{\Sigma}_i| - \frac{N_j}{2} \log |\mathbf{\Sigma}_j| - \frac{\lambda}{2}(d + \frac{d(d-1)}{2}) \cdot \log N_{ij}, \quad (5.2)$$

kjer je N_i število vzorcev v roju C_i, N_j število vzorcev v C_j, $N_{ij} = N_i + N_j$ število vzorcev v obeh rojih in Σ_{ij} ocenjena kovariančna matrika na vseh podatkih iz roja $C_i \cup C_j$. d podobno kot v primeru segmentacije predstavlja dimenzijo vektorjev značilk, s katerimi opisujemo segmente. V koraku združevanja (korak 2.2 iz sheme 5.2) iščemo tak par rojev (C_i, C_j), pri katerem dobimo minimalno vrednost kriterijske funkcije bic iz enačbe (5.2).

Pri tem moramo poudariti, da se osnovni postopek rojenja SG loči od klasične izvedbe postopkov hierarhičnega rojenja z združevanjem v tem, da se izvaja združevanje rojev s kriterijem BIC na podlagi skupnih modelov rojev, ne pa na podlagi posamičnih modelov segmentov, ki so v roju. Z drugimi besedami to pomeni, da pri vsakem združevanju dveh rojev ocenimo njun skupni model (funkcijo normalne porazdelitve), ki predstavlja osnovo za nadaljnje združevanje. Zaradi tega je potrebno voditi zgodovino združevanja za vsak roj posebej. To predstavlja pomembno konceptualno razliko v primerjavi s klasičnimi hierarhičnimi postopki rojenja z združevanjem, kjer se običajno 'razdalje' med roji ocenjujejo na podlagi razdalj med posameznimi elementi obeh rojev [Theodoridis-03, str. 455-464] (npr. z uteženo vsoto razdalj, povprečjem razdalj, minimalno ali maksimalno razdaljo med posameznimi elementi obeh rojev ipd.). Po drugi strani obstaja tudi veliko postopkov združevanja rojev na podlagi 'tipičnih' predstavnikov rojev (centroidi, variance, ipd.), kot so (po [Theodoridis-03, str. 455-464]) družina postopkov WPGMA, UPGMA, UPGMC, WPGMC ali pa Wardov algoritem, vendar se ti ne uporabljajo pri razvrščanju segmentov po govorcih. Bistvena lastnost osnovnega postopka s kriterijem BIC je ta, da razdalja med roji na vsakem koraku rojenja monotono ne narašča, ampak se spreminja glede na velikost rojev (število segmentov v roju). To pa predstavlja velik problem pri izbiri ustreznih kriterijev za zaustavitev rojenja.

Druga pomembna lastnost hierarhičnih postopkov rojenja z združevanjem je tudi ta, da ko enkrat združimo dva roja, ostaneta združena do konca rojenja. To lahko predstavlja tudi možen vir napak pri razvrščanju segmentov po govorcih, zato je potrebna pravilna izbira kriterija združevanja in skrbno načrtovanje predstavitev segmentov rojenja. V primeru osnovnega postopka rojenja SG se običajno uporabljajo akustične predstavitve segmentov, predvsem značilke koeficientov melodičnega kepstra MFCC skupaj z njihovimi odvodi ΔMFCC, ki smo jih uporabljali tudi v naših preizkusih.

Omeniti moramo še računsko in prostorsko zahtevnost osnovnega postopka rojenja SG. Na vsakem koraku t hierarhičnega rojenja z združevanjem imamo tako $N - t$ rojev, kjer predstavlja N število začetnih segmentov v posnetku. To pomeni, da moramo v vsakem koraku t pregledati $\binom{N-t}{2}$ parov rojev za sprejetje odločitve, kateri par bomo združili. Ob predpostavki, da izvedemo rojenje do konca, se pravi, da imamo N korakov rojenja, je tako število izračunov kriterija g enako $\sum_{t=0}^{N-1}\binom{N-t}{2} = \frac{(N-1)N(N+1)}{6}$. Zato je ključnega pomena za računsko zahtevnost postopka izračun kriterijske funkcije g. V primeru kriterija BIC iz enačbe (5.2) moramo za vsak par rojev (C_i, C_j) izračunati posamične in skupne ocene kovariančnih matrik in njihove determinante. Poleg tega pa moramo voditi še zgodovino združevanja segmentov v roje. Za hitrejše izračune kriterija BIC iz (5.2) tako vodimo posebne matrike vrednosti determinant in vrednosti kriterija BIC za vse možne pare rojev, ki jih na vsakem koraku postopka sproti popravljamo. To poveča prostorsko zahtevnost osnovnega postopka. V primeru rojenja velikega števila segmentov (npr. rojenja segmentov celotne zbirke) je zato bolj smiselno razdeliti dano množico segmentov na manjše podmnožice, na katerih izvedemo rojenje, potem pa še enkrat izvedemo rojenje na združenih segmentih iz vseh podmnožic.

5.3.2 Uporaba metod razpoznavanja govorcev pri rojenju z združevanjem segmentov

Drugi alternativen postopek, ki je bil prvič predstavljen v sklopu projekta *Rich Transcription* [Fiscus-05] v okviru vrednotenja postopkov razvrščanja segmentov po govorcih *(ang. "Who spoke when" evaluations)*, temelji na spoznanjih povzetih iz postopkov identifikacije in verifikacije govorcev. Postopek, ki bo opisan v nadaljevanju, je verzija osnovnega postopka, ki je bil prvič predstavljen v [Barras-04], z izboljšavami predlaganimi v [Zhu-05] in [Sinha-05].

Tudi v tem primeru gre za postopek rojenja segmentov z združevanjem, tako da osnovni koraki algoritma rojenja ostanejo enaki osnovnemu referenčnemu postopku iz sheme 5.2. Spremenijo pa se predstavitve osnovnih segmentov oziroma rojev združevanja in sam kriterij združevanja.

Ker gre pri rojenju segmentov za združevanje segmentov po govorcih, je smiselno segmente predstaviti s takšnimi predstavitvami, ki se uporabljajo pri razpoznavanju govorcev. Standarden pristop pri razpoznavanju govorcev je modeliranje govorcev z GMM modeli, [Reynolds-95]. Običajen postopek modeliranja govorcev z GMM modeli je naslednji. Najprej zgradimo splošen GMM model govora *(ang. universal background model, UBM)*, ki ga ponavadi ocenimo iz obsežnih zbirk govora. Število Gaussovih porazdelitev v UBM modelu je zato lahko zelo veliko (med 128 in 2048 Gaussovih porazdelitev). Model govora UBM se nato z različnimi postopki preoblikuje v model GMM posameznega govorca glede na njegove govorne posnetke. Najbolj pogosta transformacija UBM modela v modele GMM se izvaja na podlagi MAP adaptacije *(ang. maximum a posteriori adaptation)* [Reynolds-95], kjer povprečja posameznih Gaussovih porazdelitev modela GMM ocenjujemo iz modela UBM tako, da z modelom GMM kar najboljše opišemo podatke govorca, ki ga modeliramo. Tako imamo na eni strani z modelom UBM ocenjeno porazdelitev splošnega govora (podano s kombinacijo

normalnih porazdelitev), na drugi strani pa z modelom GMM ocenjeno porazdelitev govora posameznega govorca. Odločitev, ali dani govorni posnetek pripada določenemu govorcu, se običajno sprejme na podlagi razmerja med logaritmom vrednosti porazdelitve *(ang. log–likelihood, LLH)*, opisane z GMM , $\log(p(.|\text{GMM}))$, in logaritmom vrednosti porazdelitve iz modela UBM, $\log(p(.|\text{UBM}))$, pridobljenima iz danega posnetka. Ponavadi to razmerje označujemo z *LLR (ang. log–likelihood ratio)*. Odločitev, ali dani posnetek pripada določenemu govorcu, sprejmemo na podlagi praga odločitve. V primeru, da je *LLR* večji od praga, posnetek pripada govorcu, sicer pa ne.

Ta splošen princip delovanja sistemov za razpoznavanje govorcev pa se dá prevesti tudi na rojenje segmentov po govorcih. UBM model govora lahko dobimo na podoben način, kot je to v primeru razpoznavanja govorcev. Običajno se pri razpoznavanju govorcev uporabljajo različni UBM modeli glede na spol govorcev in/ali tip kanala govora. Ker pa v primeru razvrščanja segmentov izvajamo rojenje na danem posnetku, lahko UBM modele ocenimo kar iz posnetka, ki ga obdelujemo. V tem primeru govorimo o UBM modelih vezanih na posnetek ali krajše o UBM modelih posnetka *(ang. document speech background model)*, [Ben-04]. Ko pa enkrat imamo UBM modele, lahko vsak segment v posnetku ocenimo z ustreznim GMM modelom na podlagi prilagajanja UBM modelov s postopkom MAP adaptacije. Kriterij združevanja tako predstavljenih segmentov izpeljemo iz kriterija *LLR* in ga imenujemo *navzkrižni LLR kriterij (ang. cross log-likelihood ratio, CLR)*, [Barras-04]:

$$clr(C_i, C_j) = \frac{\log p(\mathbf{x}_i|\text{GMM}_j)}{\log p(\mathbf{x}_i|\text{UBM})} + \frac{\log p(\mathbf{x}_j|\text{GMM}_i)}{\log p(\mathbf{x}_j|\text{UBM})}. \tag{5.3}$$

C_i in C_j predstavljata segmenta oziroma roja i in j, GMM_i in GMM_j sta modela GMM ocenjena iz ustreznega UBM modela z MAP adaptacijo na podlagi vzorcev $\mathbf{x}_i$ iz roja segmentov C_i oziroma $\mathbf{x}_j$ iz roja C_j. $clr(C_i, C_j)$ tako predstavlja vsoto dveh *LLR* razmerij, s prvim preverjamo pripadnost vzorcev iz roja C_i modelu (govorcu) GMM_j, v drugem pa pripadnost roja C_j modelu GMM_i. Večja kot je vrednost $clr(C_i, C_j)$ bolj verjetno roja C_i in C_j pripadata istemu govorcu. V postopku rojenja segmentov z združevanjem iz sheme 5.2, tako v koraku 2.2 združimo tista dva roja, kjer s kriterijem *clr* dosežemo maksimalno vrednost. Pri tem moramo pri vsaki združitvi rojev ponovno naučiti GMM model združenega roja iz podatkov obeh rojev.

Uspešno delovanje takšnega rojenja je odvisno od več faktorjev. Najpoglavitnejši je ta, da moramo GMM modele zgraditi tako, da bomo z njimi dovolj dobro ocenjevali segmente posameznih govorcev. Prav zato morajo biti modeli dovolj splošni za dobro ocenjevanje kriterija CLR, hkrati pa mora učenje modelov iz UBM modelov potekati dovolj hitro za učinkovito delovanje celotnega postopka rojenja. Zato je bilo že v okviru postopkov za razpoznavanje govorcev predlaganih veliko izboljšav predvsem na področju normalizacije predstavitev govornih posnetkov, izbiri pravih modelov UBM in pri normalizaciji ocen kriterija LLR, ki so bile uspešno vključene tudi v ta postopek rojenja segmentov.

V nadaljevanju bomo opisali postopek rojenja na podlagi prilagajanja UBM modelov s postopkom MAP adaptacije (rojenje *MAP–UBM*) s predlaganimi izboljšavami iz [Barras-04, Zhu-05, Sinha-05], ki so bile prevzete iz sistemov za razpoznavanje go-

vorcev. Kot smo že omenili, gre tudi pri rojenju *MAP–UBM* za postopek rojenja z združevanjem, ki deluje po shemi 5.2. Spremembe in izboljšave osnovnega postopka lahko strnemo v naslednja področja:

- *Predstavitev segmentov:* Bistvenega pomena pri rojenju *MAP–UBM* je segmentacija, ki mora biti izvedena na način, da vsak segment predstavlja samo enega govorca, saj v nasprotnem dobimo slabe predstavitve segmentov. Predstavitve segmentov pa se izvajajo na podlagi akustične informacije. Običajno se uporabljajo značilke MFCC in njihovi odvodi ΔMFCC brez kratkočasovne energije (E), vendar z njenim odvodom ΔE. Pri tem se običajno izvaja normalizacija značilk, da se znebimo vplivov akustičnega kanala in različnih akustičnih ozadij. V našem primeru smo preizkušali dve sorodni tehniki normalizacije značilk: normalizacijo z izničevanjem skupnega povprečja in variance *(ang. cepstral mean and variance normalization, CMVN)*, [Young-04] in z metodo prileganja značilk k normalnim porazdelitvam *(ang. feature warping, FW)*, [Pelecanos-01].

- *Modeli rojenja:* Zaradi krajših odsekov govora zajetih v segmentih je potrebna pazljiva izbira števila Gaussovih porazdelitev v GMM modelih. Več kot je normalnih porazdelitev v GMM-ju, več parametrov je potrebno oceniti, po drugi strani pa krajši kot so segmenti, manj podatkov imamo za ocene. Običajno se zato izbere manjše število normalnih porazdelitev v UBM modelih, ki jih potem s postopkom MAP adaptacije prevedemo v GMM modele segmentov. Če so modeli UBM naučeni na splošnih zbirkah govora, se običajno pri rojenju *MAP–UBM* uporablja 128 normalnih porazdelitev, [Barras-04, Sinha-05]. V našem primeru pa smo uporabljali UBM modele pridobljene na trenutnih posnetkih, zato smo se omejili samo na 16 gostot porazdelitev.

  Drugo vprašanje je, kakšne in koliko UBM modelov uporabljamo pri rojenju. Običajno se uporablja različne UBM modele glede na spol govorcev in na akustične kanale govora (čist govor, telefonski govor ipd.) V primeru našega rojenja smo uporabljali ločene UBM modele za ženske in moške govorce. Pri tem smo morali izvesti dodatno klasifikacijo segmentov glede na spol.

  Zaradi splošnosti modelov in hitrosti rojenja pa je potrebno tudi prilagoditi parametre in omejiti število korakov postopka MAP adaptacije. Ponavadi je število korakov adaptacije omejeno na manj kot 5. V [Sinha-05] so preizkusili številne kombinacije postopka MAP adaptacije in izkazalo se je, da je izbira dveh korakov v postopku adaptacije dober kompromis med natančnostjo in učinkovitostjo rojenja. Tudi v naših preizkusih smo zato uporabljali dva koraka učenja GMM modelov ob optimalni izbiri ostalih parametrov postopka MAP adaptacije.

- *Združevanje rojev:* Za kriterij združevanja rojev segmentov se uporablja kriterij *clr* iz enačbe (5.3). Na vsakem koraku združevanja poiščemo tisti par rojev (C_i, C_j), pri katerem dosežemo maksimalno vrednost kriterija *clr*. Za vsak tako združeni roj moramo dodatno pridobiti GMM model s pomočjo prilagajanja UBM modela s postopkom MAP adaptacije. V primeru več UBM modelov združujemo samo tiste roje, pri katerih so GMM modeli naučeni iz istega UBM modela. V našem primeru to pomeni, da ločeno rojimo segmente, ki pripadajo različnim spolom.

Poleg kriterija CLR smo preizkušali tudi druge kriterije, s katerimi smo bistveno presegli rezultate razvrščanja, in bodo predstavljeni v naslednjem razdelku.

Časovna in prostorska zahtevnost opisanega postopka rojenja je bistveno večja kot pri osnovnem referenčnem postopku in je predvsem odvisna od števila normalnih porazdelitev GMM modelov in od postopka MAP adaptacije. V prvem koraku postopka rojenja (korak 1 iz sheme 5.2) moramo tako za vsak segment pridobiti po en GMM model. Potem pa moramo na vsakem koraku združevanja oceniti še po en GMM model združenega roja. Tako moramo v primeru, ko imamo N začetnih segmentov, izvesti najmanj $2N$ MAP adaptacij, zato je bistvenega pomena za hitrost delovanja postopka rojenja *MAP–UBM* ravno izvajanje MAP adaptacij. Tu moramo zato vedno tehtati med hitrostjo konvergence parametrov UBM modela k želenemu GMM modelu in kvaliteto ocenjenega modela. Tako poskušamo v čim manj korakih postopka MAP adaptacije pridobiti čim boljše ocene modelov. Seveda je potrebno pridobiti tudi ustrezne UBM modele. Če jih učimo na splošnih zbirkah govora, jih pripravimo že vnaprej, če pa uporabljamo UBM modele posnetkov, jih zgradimo v prvi fazi postopka rojenja.

V nadaljevanju bomo predstavili še dva kriterija združevanja rojev na podlagi *MAP–UBM* predstavitev segmentov, s katerimi smo znatno izboljšali rezultate razvrščanja segmentov v primerjavi z osnovnim kriterijem CLR.

5.3.2.1 Predlagani kriterij združevanja segmentov

V postopku rojenja *MAP–UBM* je osnovni kriterij združevanja rojev *clr* iz enačbe (5.3), sestavljen iz vsote dveh razmerij *LLR*, kjer v obeh primerih navzkrižno ugotavljamo, ali je bolj verjetno, da je zaporedje vzorcev enega roja porodil naključni proces, opisan z GMM modelom drugega roja, ali da jih je tvoril UBM model. Če je ujemanje vzorcev z GMM modelom v obeh primerih večje od ujemanja z UBM modelom, je vrednost *clr* velika in sklepamo lahko, da roja pripadata istemu skupnemu viru oziroma istemu govorcu. Pri tem pa igra bistveno vlogo model UBM. 'Dober' UBM model zagotavlja zanesljivejše ocene in s tem učinkovito delovanje kriterija *clr*, v primeru slabega UBM modela in slabih ocen GMM modelov (kar je v primeru kratkih segmentov pogosto) pa se na podlagi *clr* kriterija ne moremo zanesljivo odločati med dvema možnostima. Zato smo hoteli zmanjšati neposreden vpliv UBM modela na delovanje kriterija. To smo naredili tako, da smo iskali takšne kriterije oziroma mere podobnosti, pri katerih se izvaja primerjava samo na predstavitvah GMM modelov in UBM modelov neposredno ne vključujejo. Kljub temu pa ostanejo UBM modeli posredno vključeni v GMM modelih, saj so GMM modeli izpeljani s postopki prilagajanja iz UBM modelov (MAP adaptacija).

Takšnih verjetnostnih mer je več [Pavešić-00, str. 168], [Theodoridis-03, poglavje 5], vendar imamo v našem primeru pri vseh enak problem. Medtem ko so v primeru ene normalne porazdelitve možne analitične izpeljave posameznih verjetnostnih mer, pa v primeru modeliranja porazdelitev z večjim številom Gaussovih porazdelitev analitičnih rešitev ni. Pridobivanje ocen tako poteka neposredno iz definicij verjetnostnih mer z različnimi metodami (najpogosteje z metodami Monte Carlo, npr. [Ben-02]), ki pa so računsko zelo zahtevne in jih v praksi zato ne izvajamo. Izkazalo se je, da

se dá v primeru GMM modelov pridobljenih s postopkom MAP adaptacije, kjer poteka prilagajanje samo povprečnih vrednosti v Gaussovih porazdelitvah UBM modelov, učinkovito oceniti zgornjo mejo Kullback-Leiblerjeve mere [Do-03, Ramos-Castro-05]. Po zgledu iz [Ben-04] smo zato vpeljali dodaten kriterij združevanja rojev na podlagi tako pridobljenih GMM modelov:

$$upkl2(C_i, C_j) = \sqrt{\sum_{k=1}^{K}\sum_{d=1}^{D} w_k \cdot \frac{(m_{k,d}^{C_i} - m_{k,d}^{C_j})^2}{\sigma_{k,d}^2}}, \tag{5.4}$$

kjer so $m_{k,d}^{C_i}$ in $m_{k,d}^{C_j}$ povprečne vrednosti iz k-tega povprečnega vektorja na d-tem mestu GMM modela roja C_i oziroma roja C_j. w_k in $\sigma_{k,d}^2$ pa so ustrezne uteži in variance UBM modela, iz katerega ocenjujemo GMM modele rojev C_i in C_j. Pri tem predpostavimo, da imamo K normalnih porazdelitev v UBM modelu in dimenzijo značilk D. Kriterij $upkl2$ predstavlja zgornjo mejo za oceno simetrične divergenčne KL2 mere [Pavešić-00, str.168]. Manjše vrednosti $upkl2(C_i, C_j)$ pomenijo večjo podobnost porazdelitev GMM modelov rojev oziroma večjo podobnost rojev C_i in C_j. V postopku rojenja segmentov z združevanjem iz sheme 5.2 na vsakem koraku rojenja tako združujemo tiste roje, kjer dosežemo s kriterijem $upkl2$ najmanjšo vrednost.

Drugi kriterij, s katerim smo v naših eksperimentih dosegali še boljše rezultate razvrščanja segmentov, je bil izveden na podlagi kriterija BIC, ki ga uporabljajmo tudi v osnovnem postopku rojenja z združevanjem. Osnovna verzija kriterija BIC iz enačbe (4.3), ki je bila predstavljena v poglavju segmentacije zvočnih posnetkov, pravilno deluje samo v primeru, ko je vsak model opisan samo z eno Gaussovo porazdelitvijo. Vendar pa se je v številnih praktičnih aplikacijah [Campbell-97, Dasgupta-98, Leroux-92, Roeder-97] izkazal za primerno mero podobnosti tudi med modeli opisanimi z večjim številom Gaussovih porazdelitev. Kriterij BIC iz enačbe (4.3) tako prevedemo v primeru rojenja z GMM modeli na:

$$\begin{aligned} bic(C_i, C_j) \;=\; & \sum_{n=1}^{N_i} \log p(\mathbf{x}_n\,|\mathrm{GMM}_i) + \sum_{n=1}^{N_j} \log p(\mathbf{x}_n\,|\mathrm{GMM}_j) - \sum_{n=1}^{N_i+N_j} \log p(\mathbf{x}_n\,|\mathrm{GMM}_{ij}) \\ & -\frac{\lambda}{2} \cdot (K \cdot (2D+1)) \cdot \log(N_i + N_j). \end{aligned} \tag{5.5}$$

Ta izvedba kriterija BIC je enaka osnovni verziji iz enačbe (5.2), le da eno normalno porazdelitev za vsak segment (roj) nadomestimo s K porazdelitvami iz GMM modelov rojev C_i, C_j in skupnega združenega roja $C_i \cup C_j$. Pri tem računamo v vseh primerih logaritme vrednosti porazdelitev opisane z modeli GMM_i, GMM_j in GMM_{ij} za vse podatke iz ustreznih rojev. V drugem delu kriterija bic z utežnim faktorjem λ uravnavamo razmerje med kvaliteto modelov GMM in njihovo kompleksnostjo, ki jo merimo s produktom $(K \cdot (2D+1)) \cdot \log(N_i + N_j)$, kjer predstavlja K število Gaussovih porazdelitev v modelu GMM, D dimenzijo vektorja značilk in $N_i + N_j$ število vseh vzorcev, ki jih obdelujemo.

V obeh kriterijih UBM model ni neposredno vključen, ampak je le osnova za oceno povprečij GMM modelov posameznih rojev oziroma segmentov. Razlika med $upkl2$ iz (5.4) in kriterijem bic iz (5.5) je predvsem v tem, da v primeru $upkl2$ primerjamo

dve porazdelitvi rojev neposredno, v primeru *bic* pa posredno preko skupnega modela združenega roja GMM_{ij}. Na ta način izvajamo normalizacijo ocen logaritmov vrednosti porazdelitev preko skupnega modela porazdelitev, kar je na nek način primerljivo z osnovnim *clr* kriterijem iz enačbe (5.3), kjer pa ocene na drugačen način normaliziramo preko globalnih UBM modelov.

V nadaljevanju bomo pokazali, da je izbira kriterijev postopkov rojenja *MAP–UBM* bistveno vplivala na rezultate razvrščanja segmentov.

5.4 Postopek rojenja segmentov z združevanjem akustične in prozodične informacije

V tem razdelku bomo opisali izvirni postopek rojenja segmentov z združevanjem na podlagi dveh različnih predstavitev govornih odsekov. Pri tem smo za osnovni postopek rojenja segmentov uporabili rojenje z združevanjem rojev opisanem iz sheme 5.2. Razlika pa je bila v predstavitvah govornih odsekov, ki jih rojimo. Pri tem smo se zgledovali po najbolj uspešnih metodah razpoznavanja govorcev, ki temeljijo na združevanju več nivojev predstavitev govorčevih lastnosti in njegovega govora [Kajarekar-03, Shriberg-05, Reynolds-03a, Baker-05]. V tem primeru pri razpoznavanju govorcev poleg osnovnih značilnosti govorcev opisujemo tudi način govora posameznega govorca, tipične jezikovne lastnosti govorca, posebnosti v izgovorjavi ipd. Z dodatno informacijo tako pridobimo dvoje: nove značilnosti posameznih govorcev, ki jih uporabljamo kot dopolnilno informacijo osnovnim akustičnim predstavitvam govorcev, in zaradi načina pridobivanja teh značilk večjo neobčutljivost na akustične spremembe v govornih posnetkih, kar posledično pomeni večjo zanesljivost pri odločanju o posameznih govorcih. Tako so se sistemi, kjer odločanje o govorcih temelji na združevanju osnovnih akustičnih lastnosti govorca s prozodično in jezikovno informacijo, izkazali za najboljše med sistemi za razpoznavanje govorcev.

Tudi v našem primeru razvrščanja segmentov po govorcih smo zato hoteli preizkusiti postopke rojenja, ki bi temeljili na združevanju akustične in prozodične informacije govornih odsekov. Pri tem je potrebno poudariti, da je v primeru rojenja segmentov zvočnih posnetkov informativnih oddaj poleg akustične informacije smiselno uporabljati tudi prozodično informacijo, saj se način govora in razpoloženje govorcev znotraj ene informativne oddaje bistveno ne spreminja.

V nadaljevanju bomo opisali, kako smo pridobivali značilke na podlagi akustičnih in prozodičnih lastnosti govornih odsekov. Predstavili bomo način integracije obeh informacij v primeru rojenja segmentov z združevanjem. Pri tem se bomo ukvarjali predvsem s kriteriji združevanja rojev na podlagi obeh informacij, z normiranjem ocen kriterijev in s postopki odločanja na podlagi uteženih ocen obeh predstavitev. Osnovno vodilo pri združevanju je bilo tudi v našem primeru, da prozodične predstavitve predstavljajo le dopolnilno informacijo pri rojenju v primeru, ko odločitve o združevanju rojev samo na podlagi akustičnih predstavitev ne moremo sprejeti.

5.4.1 Pridobivanje akustičnih lastnosti govornih odsekov

Predstavitve, ki temeljijo na akustičnih lastnostih govornih odsekov, ponavadi predstavimo z zaporedjem vektorjev značilk, ki jih tvorimo iz govornega signala na vsakih nekaj milisekund. Zato lahko govorimo, da z njimi opisujemo kratkočasovne lastnosti danih govornih odsekov. V primeru razpoznavanja govorcev ali pa v našem primeru pri rojenju segmentov nam to zaporedje predstavlja osnovo za oceno modelov ali verjetnostnih porazdelitev, s katerimi želimo opisati naključne vire, ki so porodili dana zaporedja vektorjev značilk. Odločitev, ali dva odseka govornega signala pripadata istemu viru, sprejmemo na podlagi primerjav med temi modeli oziroma njihovimi verjetnostnimi porazdelitvami. Tako lahko akustične predstavitve govornih odsekov obravnavamo kot kombinacijo akustične značilke – model predstavitve, kjer na podlagi zaporedja kratkočasovnih akustičnih dogodkov z modeli ocenjujemo globalne akustične lastnosti govornih odsekov.

To je seveda običajna praksa modeliranja akustičnih lastnosti govorcev v sistemih za razpoznavanje govorcev in smo jo že uporabili za modeliranje govornih odsekov pri segmentaciji in pri osnovnih postopkih rojenja zvočnih posnetkov. Gre pa za pomembno konceptualno razliko v primerjavi s prozodičnimi predstavitvami, kjer značilke izpeljemo iz osnovnih predstavitev govora in jih ocenjujemo na daljših odsekih govornih signalov na podlagi drugačnih osnovnih enot.

V našem primeru smo za akustične značilke izbrali standardne značilke koeficientov melodičnega kepstra MFCC in njihove odvode ΔMFCC, ki smo jih uporabljali tudi v osnovnem postopku rojenja. Govorni odseki pa so bili podobno kot v referenčnem postopku modelirani z eno Gaussovo porazdelitvijo.

5.4.2 Pridobivanje prozodičnih lastnosti govornih odsekov

Kot smo že omenili, smo prozodične značilke uporabljali kot dodatno informacijo pri odločitvah o združevanju rojev. To je smiselno zaradi dveh razlogov. Prozodične značilke običajno izpeljemo na podlagi osnovnih predstavitev govora, zato jih ocenjujemo na daljših odsekih govornih signalov. Za zanesljive ocene potrebujemo več govornih podatkov, kar v našem primeru pomeni, da daljši kot so govorni odseki (večji segmenti, roji), boljše bodo ocene prozodičnih značilk, in nasprotno. Drugi razlog pa je v lastnostih, ki jih opisujemo s prozodičnimi značilkami. Z njimi namreč opisujemo bolj način in stil govora zajetega v danih govornih odsekih, kar pomeni, da bi v primeru rojenja samo na podlagi prozodičnih značilk, združevali bolj enake tipe govora kot pa same govorce. To pa seveda ni bil naš cilj.

Pri načrtovanju značilk smo se zgledovali predvsem po izvedbi prozodičnih značilk za razpoznavanje govorcev predstavljenih v [Shriberg-05]. V splošnem prozodične značilke tvorimo iz treh osnovnih predstavitev govornih signalov [Nöth-02]: energije (glasnosti) signala *(eng)*, višine osnovnega tona govora *(ang. pitch, f_0)* in osnovnih enot govora, ki jih običajno predstavimo z besedami. Na podlagi teh predstavitev izpeljemo tri skupine prozodičnih značilk: *energijske značilke (ang. energy features)*, *značilke trajanja (ang. duration features)* in *značilke f_0 (ang. pitch features)*. Izvedba prozodičnih značilk je

odvisna od namenov uporabe in različnih pravil izpeljave. Običajen postopek je, da izpeljemo čim več različnih značilk (od nekaj 10 do nekaj 100), ki jih potem z različnimi tehnikami redukcije dimenzije značilk (PCA, LDA ipd.) zmanjšamo na željeno število, primerno za nadaljnjo obdelavo.

V primeru razpoznavanja govorcev je potrebno osnovne enote govora za izvedbo prozodičnih značilk zmanjšati. V primeru sistema za razpoznavanje govorcev [Shriberg-05] so namesto besed za osnovne enote govora uporabljali zloge razpoznanega govora. V našem primeru pa smo se odločili za uporabo zvenečih *(ang. voiced, V)* in nezvenečih *(ang. unvoiced, U)* odsekov govora kot osnovnih govornih enot za tvorbo prozodičnih značilk.

Postopek izvedbe prozodičnih značilk je bil naslednji. Najprej smo ocenili glasnost (energijo) signala na podlagi logaritma vsote kvadratov amplitud signala. Pri tem nismo izvajali normiranja energije. Potem smo s postopkom *ESPS/Waves* [Talkin-95] izračunali potek višine osnovnega tona $f0$ na celotnem signalu. Vrednosti f_0 smo še dodatno filtrirali z medianinim filtrom, da bi se znebili potencialnih napak iz osnovnega postopka. Tako energija kot f_0 sta bila ocenjena na vsakih 10 ms govornega signala. Za ocenjevanje zvenečih in nezvenečih področij govornih signalov smo uporabljali razpoznavalnik glasov slovenskega govora. Uporabljeni razpoznavalnik je bil isti, kot smo ga že uporabljali pri izvedbi značilk za detekcijo govornih in ne–govornih delov in je bil podrobneje opisan v poglavju 3. Rezultat razpoznavanja so bile transkripcije govornih signalov po glasovih skupaj s trajanjem posameznih glasov. Glasove smo pretvorili v zveneče (U) in nezveneče (V) simbole ter na simbole, ki so predstavljali premore v govornih posnetkih (S). Pred izvedbo prozodičnih značilk smo jih dodatno uskladili s potekom f_0. Tako smo govorni signal opisali s tremi predstavitvami: energijo ali glasnostjo, potekom f_0 ter s potekom in trajanjem osnovnih govornih enot V, U, S.

Na podlagi teh predstavitev smo na vsakem govornem segmentu tvorili 10 prozodičnih značilk, ki jih lahko razdelimo na tri skupine glede na izvedbo iz osnovnih predstavitev:

1. **Energijske značilke:**
   - *povprečna energija* je povprečna vrednost kratkočasovne energije ocenjena na govornem odseku;
   - *varianca energije* je ocenjena varianca kratkočasovne energije na govornem odseku;
   - *število naraščajočih točk energije* predstavlja razmerje med številom točk naraščanja energije govornega signala in vsemi točkami izračuna energije na celotnem govornem odseku;
   - *število padajočih točk energije* predstavlja razmerje med številom točk padanja energije govornega signala in vsemi točkami izračuna energije na celotnem govornem odseku.
2. **Značilke trajanja:**
   - *hitrost govora* na podlagi V, U, S enot, kjer štejemo število zamenjav osnovnih enot V, U, S na govornem segmentu in jih delimo s trajanjem govornega segmenta;

- *razmerje med povprečnim trajanjem V in U enot* računamo podobno kot pri GNG segmentaciji po formuli (3.4). Predstavlja razmerje med povprečnim trajanjem zvenečih (V) in nezvenečih glasov (U) normirano s trajanjem govornega odseka, kjer računamo obe povprečji.

3. **Značilke f_0:**

- *povprečen* f_0 je povprečna vrednost f_0 izračunana samo na govornih odsekih zvenečih glasov;
- *varianca* f_0 je ocenjena varianca poteka f_0 izračunana prav tako samo na govornih odsekih zvenečih glasov;
- *število naraščajočih točk* f_0 podobno kot v primeru energije predstavlja razmerje med številom naraščajočih točk f_0 v primerjavi z vsemi točkami izračuna f_0 (brez U in S enot) na celotnem govornem odseku;
- *število padajočih točk* f_0 pa predstavlja razmerje med številom padajočih točk f_0 v primerjavi z vsemi točkami izračuna f_0 na govornem odseku.

Pri načrtovanju izbranih prozodičnih značilk smo se zgledovali po izvedbi prozodičnih značilk za razpoznavanje govorcev [Shriberg-05] in po značilkah za detekcijo prozodičnih mej [Gallwitz-02]. Razlika je bila predvsem v tem, da smo v našem primeru izbrali za osnovne enote izračunov značilk enote VUS oziroma sklenjena področja zvenečih glasov, ki smo jih predhodno pridobili z ustreznim razpoznavalnikom glasov. V primeru prozodičnih značilk za razpoznavanje govorcev [Shriberg-05] so za osnovne enote uporabljali zloge govora, v primeru [Nöth-02] pa so značilke tvorili na podlagi samodejno pridobljenih besednih prepisov govora. V našem primeru smo morali prilagoditi izračune prozodičnih značilk na sorazmerno kratke govorne segmente, zato smo tudi izbirali takšne značilke, ki bi jih lahko dovolj dobro ocenili iz danih segmentov. Po drugi strani smo morali za značilke izbirati tudi takšne prozodične lastnosti, ki bi vsebovale kar največ informacije za ločevanje med govorci. Med različnimi izvedbami značilk iz [Nöth-02, Shriberg-05] smo izbrali takšne, ki jih lahko ocenjujemo tudi iz krajših govornih odsekov in še nosijo dovolj informacije o govorcih.

Druga pomembna zahteva pri načrtovanju prozodičnih značilk je bilo zagotavljanje neobčutljivosti značilk na akustične spremembe v govoru in na spremenljive dolžine govornih odsekov. Vse značilke ustrezno je bilo zato potrebno normirati. Izvajali smo dva tipa normiranja: povprečenje količin in normiranje po času glede na trajanje oziroma število točk v govornem odseku. Poseben primer so bile značilke f_0, kjer smo izpeljali značilke samo iz zvenečih glasov in zato je bilo potrebno značilke normirati glede na število zvenečih delov v danem govornem odseku. S tem smo dosegli ustrezno primerljivost različno dolgih govornih odsekov in zmanjšali občutljivost na napake iz osnovnih prozodičnih predstavitev.

Predlaganih 10 prozodičnih značilk se je izkazalo v kombinaciji z akustičnimi značilkami za dovolj robustne in zanesljive pri rojenju z združevanjem segmentov po govorcih.

5.4.3 Predlagani postopek rojenja segmentov

Osnovni namen izvedbe postopka rojenja z združevanjem govornih odsekov predstavljenih z dvema predstavitvama je bil v tem, da smo hoteli izboljšati delovanje postopka rojenja segmentov istih govorcev ob spremenjenih akustičnih pogojih. Pri standardnem postopku rojenja z združevanjem samo na podlagi akustičnih značilk se je namreč pokazalo, da se na neki stopnji rojenja bolje združujejo roji, ki imajo podobne akustične lastnosti (enako akustično ozadje, enake pogoje zajema govornih posnetkov), kot pa roji, ki vsebujejo govorne odseke istih govorcev. To je sicer pričakovano in celo priporočljivo, če uporabljamo takšno razvrščanje segmentov za namene prilagajanja modelov po govorcih. Naš cilj pa je bil razvrščanje segmentov po govorcih, zato smo poleg akustične informacije hoteli vključiti še prozodično, ki je manj občutljiva na akustične spremembe govornih odsekov. Pri tem smo sledili osnovnemu principu, da bi prozodično informacijo uporabljali kot dopolnilno k osnovni akustični informaciji. Tako je bil predlagani postopek rojenja načrtovan na način, da bi na začetku rojenja, ko imamo na razpolago manj govornih podatkov, združevali roje bolj na podlagi akustične informacije, daljše govorne odseke oziroma roje z dovolj govornih podatkov pa bi združevali bolj na podlagi prozodične informacije. Osnovno vodilo pri tem je bilo, da bi najprej samo na podlagi akustične informacije združevali govorne odseke, ki so akustično podobni, torej segmente, ki vključujejo istega govorca ob enakih akustičnih pogojih, z dodatno prozodično informacijo pa bi se v nadaljevanju usmerili bolj proti združevanju na podlagi enakih govornih lastnosti govorcev.

Postopek rojenja z akustičnimi in prozodičnimi značilkami govornih odsekov je bil zasnovan na podlagi postopka rojenja z združevanjem predstavljenim s shemo 5.2. Pri tem smo spremenili predstavitve govornih odsekov v rojih in kriterij združevanja rojev.

Vsak govorni odsek je bil predstavljen z akustičnimi in prozodičnimi značilkami. Pri akustični predstavitvi smo uporabili MFCC in ΔMFCC značilke, ki smo jih na vsakem segmentu modelirali z eno Gaussovo porazdelitvijo, torej s povprečnim vektorjem $\mathbf{m}$ in (polno) kovariančno matriko $\mathbf{\Sigma}$. V primeru prozodične predstavitve pa smo vsak segment predstavili z vektorjem desetih značilk, ki smo jih opisali v prejšnjem razdelku.

Združevanje govornih odsekov oziroma rojev, ki so vključevali govorne odseke, je potekalo s kombinacijo ocen združevanj na podlagi obeh predstavitev. Za kriterij združevanja na podlagi akustičnih predstavitev smo uporabili kriterij BIC iz enačbe (5.2). Pri tem je bilo potrebno oceniti λ, ki smo jo izbrali na podlagi optimalnega rojenja segmentov iz osnovnega postopka na razvojni zbirki. Poudariti moramo še, da pri kriteriju BIC združujemo tiste roje, kjer je vrednost kriterija minimalna.

Kriterij združevanja segmentov na podlagi prozodičnih značilk je bil izpeljan z uporabo razdalje v prostoru osnovnih komponent prozodičnih značilk, ki smo ga ocenili na vsakem posnetku posebej. Postopek izpeljave kriterija združevanja na vsakem posnetku je bil naslednji:

1. Na vsakem segmentu s_i smo ocenili vektor $\mathbf{v}_{s_i}^{proz}$ sestavljen iz 10-ih prozodičnih značilk.

2. Izvedli smo analizo osnovnih komponent PCA, [Pavešić-00, str. 192-197], na

vektorjih $\mathbf{v}_{s_i}^{proz}$. Pri tem smo v PCA postopku namesto kovariančne matrike uporabljali korelacijsko matriko $\mathbf{R}^{proz}$, da smo se znebili vpliva različnih številskih področij vrednosti meritev prozodičnih značilk. Tako smo dobili znani razcep $\mathbf{R}^{proz} = \mathbf{P} \cdot \mathbf{\Lambda} \cdot \mathbf{P}^T$, kjer je $\mathbf{P}$ matrika lastnih vektorjev in $\mathbf{\Lambda}$ diagonalna matrika lastnih vrednosti. Če iz matrike $\mathbf{\Lambda}$ izberemo $1 \leq L \leq 10$ lastnih vrednosti, ki so urejene od največje do najmanjše, potem označimo s $\mathbf{P}_L$ matriko lastnih vektorjev, ki ustrezajo prvim (največjim) L lastnim vrednostim $\lambda_n,\ n = 1, \ldots L$.

3. Izračunamo Mahalanobisovo razdaljo med segmenti v prostoru osnovnih komponent:

$$d_{proz}(s_i, s_j) = \sum_{n=1}^{L} \frac{(w_{s_i}^n - w_{s_j}^n)^2}{\lambda_n}, \tag{5.6}$$

kjer so $w_{s_i}^n$ komponente vektorja $\mathbf{w}_{s_i} = \mathbf{P}_L^T \cdot \mathbf{v}_{s_i}^{proz}$ pri $\lambda_n,\ n = 1, \ldots L$. Podobno velja tudi za $w_{s_j}^n$ iz segmenta s_j.

4. Mera podobnosti med rojem C_i sestavljenim iz segmentov $\{s_i \,|\, i = 1, \ldots N_i\}$ in rojem C_j sestavljenim iz segmentov $\{s_j \,|\, j = 1, \ldots N_j\}$ izračunamo kot povprečje razdalj d_{proz} po vseh parih segmentov iz rojev C_i in C_j:

$$proz(C_i, C_j) = \frac{1}{N_i N_j} \sum_{s_i \in C_i} \sum_{s_j \in C_j} d_{proz}(s_i, s_j). \tag{5.7}$$

Manjši kot je $proz(C_i, C_j)$, večja je podobnost med rojema C_i in C_j na podlagi prozodičnih značilk.

Pri izpeljavi kriterija *proz* iz (5.7) smo se zgledovali po osnovnih kriterijih rojenja s primerjavo vseh razdalj med elementi iz dveh rojev. Za osnovno razdaljo smo uporabili Mahalanobisovo razdaljo. Posebnost kriterija je uporaba PCA analize na segmentih ločeno na vsakem zvočnem posnetku. S tem smo hoteli dodatno povečati razlikovanje med segmenti in odpraviti morebitne soodvisnosti med prozodičnimi značilkami. Izbira števila lastnih vektorjev osnovnih PCA komponent je bila izvedena na razvojni zbirki in bo predstavljena v naslednjem razdelku.

Na podlagi vrednosti ocen obeh predstavitev smo se morali v vsakem koraku postopka rojenja odločati, katere roje je potrebno združevati. Pri tem smo skupni kriterij združevanja sestavili iz utežene vsote normiranih ocen obeh kriterijev. Zato v tem primeru govorimo o postopku rojenja z združevanjem segmentov s fuzijo vrednosti ocen obeh predstavitev *(ang. score–based fusion)*. Da bi lahko kontrolirali vlogo obeh predstavitev, smo morali ustrezno normirati ocene kriterijev predstavitev. Postopek normiranja je bil izveden na podlagi *min–max normalizacije* [Jain-05], ki smo jo že predstavili v poglavju 4 pri izvedbi segmentacije z združevanjem. Razlogi za uporabo min–max normalizacije so bili podobni kot pri segmentaciji. Ker se v primeru obeh kriterijev odločamo za združevanje na podlagi minimalnih vrednosti obeh kriterijev, smo morali to ohraniti tudi po normalizaciji. Ocene minimalnih in maksimalnih vrednosti obeh kriterijev potrebne za normalizacijo smo pridobili iz vrednosti osnovnih kriterijev, ki smo jih morali izračunati na vsakem koraku rojenja.

Tako se je postopek rojenja s fuzijo akustičnih in prozodičnih predstavitev segmentov razlikoval od osnovnega postopka rojenja z združevanjem iz sheme 5.2 v koraku združevanja rojev (korak 2.2 iz sheme 5.2), kjer je bilo potrebno za vsak roj predstavljen z dvema predstavitvama izvesti normalizacijo vrednosti kriterijev obeh predstavitev in nato fuzijo ocen. Korak 2.2 iz sheme 5.2 se je zato v našem primeru spremenil v:

– 2.2. Med vsemi pari rojev (C_r, C_s) v $\mathcal{S}_{t-1}$ najdemo tak par (C_i, C_j), za katerega velja:

$$fuz(C_i, C_j) = \arg\min_{C_r, C_s} fuz(C_r, C_s),$$

pri čemer je *fuz* kriterij združevanja s fuzijo:

$$fuz(C_r, C_s) = fw \cdot norm\_bic(C_r, C_s) + (1 - fw) \cdot norm\_proz(C_r, C_s). \quad (5.8)$$

Kriterij $norm\_bic(C_r, C_s)$ je normirana verzija kriterija BIC iz enačbe (5.2) in predstavlja kriterij združevanja rojev (C_r, C_s) na podlagi akustičnih predstavitev. Kriterij $norm\_proz(C_r, C_s)$ je normirani kriterij iz enačbe(5.7) in predstavlja mero podobnosti rojev na podlagi prozodičnih značilk. fw je utež fuzije.

Pri obeh kriterijih smo uporabili *min–max normalizacijo*, torej:

$$norm\_bic(C_r, C_s) = \frac{bic(C_r, C_s) - mn_{bic}(t)}{mx_{bic}(t) - mn_{bic}(t)},$$

kjer je $mn_{bic}(t)$ minimalna in $mx_{bic}(t)$ maksimalna vrednost kriterija *bic* med vsemi roji na koraku združevanja t. Podobno je tudi

$$norm\_proz(C_r, C_s) = \frac{proz(C_r, C_s) - mn_{proz}(t)}{mx_{proz}(t) - mn_{proz}(t)},$$

kjer je $mn_{proz}(t)$ minimalna in $mx_{proz}(t)$ maksimalna vrednost kriterija *proz* med vsemi roji na koraku združevanja t.

Dodatni vhodni parametri postopka so λ pri kriteriju *bic*, število lastnih vektorjev pri kriteriju za prozodijo, *proz*, in utež fuzije fw, ki smo jih določali na podlagi optimalnih rezultatov razvrščanja segmentov na razvojni zbirki.

5.5 Preizkusi postopkov rojenja

Preizkušali smo tri postopke rojenja: osnovni postopek rojenja z združevanjem segmentov na podlagi kriterija BIC, postopek rojenja *MAP–UBM* z uporabo kriterijev CLR in BIC ter postopek rojenja z združevanjem akustične in prozodične informacije.

Primerjava postopkov je bila izvedena na zbirkah SiBN in COST278. Podobno kot pri segmentaciji smo del zbirke SiBN uporabili za razvojno zbirko, s katero smo ocenjevali odprte parametre postopkov. Preostali del zbirke SiBN je bil uporabljen za testni del. Dodatno smo preizkušali postopke v primeru idealnih in samodejnih segmentacij. Pri idealnih segmentacijah smo segmente za rojenje pridobili iz ročno označenih posnetkov.

V primeru samodejnih segmentacij pa smo izvedli segmentacijo z najboljšim izmed postopkov segmentacij, ki smo jih preizkušali v prejšnjem poglavju. Dodatno smo izvedli še razvrščanje segmentov na govorne in ne–govorne dele s predlaganim postopkom, opisanim v poglavju 3.

Primerjava postopkov je bila izvedena brez *kriterija zaustavitve* rojenja, s katerim smo se posebej ukvarjali in bo opisan v nadaljevanju. S tem smo hoteli doseči neposredno primerjavo postopkov rojenja samo na podlagi različnih predstavitev segmentov in različnih kriterijev združevanja.

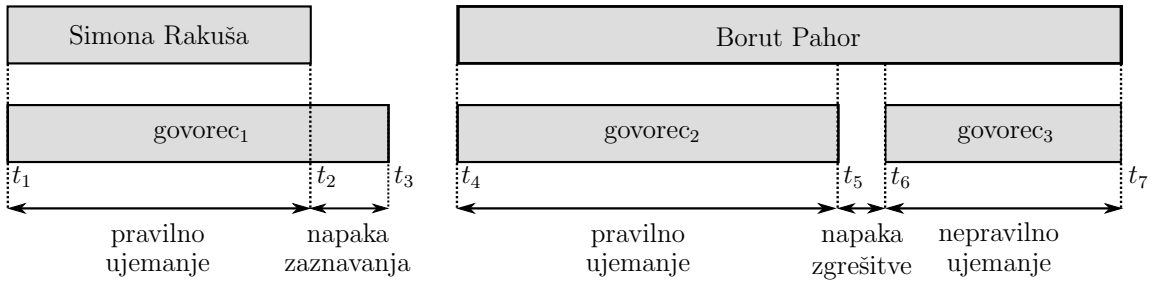
5.5.1 Vrednotenje postopkov rojenja

Za vrednotenje postopkov razvrščanja segmentov po govorcih se uporabljajo različne metode merjenja napak, ki so odvisne od nadaljnje uporabe tako strukturiranih posnetkov.

Če bi uporabljali razvrščanje segmentov za namene prilagajanja modelov po govorcih, potem je smiselno meriti kvaliteto čistosti rojev (*ang. purity*), ki jih uporabljamo potem pri postopku prilagajanja. To pomeni, da na eni strani merimo, kako dobro z enim rojem opišemo določenega govorca (*ang. cluster purity*), po drugi pa kako dobro je en govorec opisan z določenim rojem (*ang. speaker purity*). Kombinacija obeh ocen se imenuje *mera Q* (*ang. Q–measure*) in je bila predstavljena v [Ajmera-04].

Druga mera, s katero merimo celotno strukturiranje posnetkov glede na njihove dejanske (referenčne) oznake, je *mera DER* (*ang. diarisation error rate, DER*). Z njo poleg ujemanja rojev z referenčnimi govorci merimo še napake segmentacije in napake detekcije ne–govornih delov. Metoda DER je bila razvita v sklopu projekta *Rich Transcription* [Fiscus-05] v okviru vrednotenja postopkov razvrščanja segmentov po govorcih *(ang. "Who spoke when" evaluations)* in se skoraj izključno uporablja za vrednotenje postopkov rojenja segmentov v primeru strukturiranja posnetkov za namene samodejne indeksacije in iskanja govorcev v zbirkah zvočnih posnetkov.

Zato smo v naših preizkusih za vrednotenje postopkov rojenja uporabili *mero DER*, ki jo bomo predstavili v nadaljevanju.



Slika 5.3: Merjenje napak razvrščanja segmentov glede na referenčne oznake z mero DER.

Z *mero DER* ocenjujemo tri vrste napak: napako ujemanja med roji in referenčnimi govorci (*ang. speaker error*), napako zaznavanja (*ang. false alarm*) in napako zgrešitve

(*ang. miss error*). Vse tri vrste napak so prikazane na sliki 5.3. Napake merimo s časom trajanja napačno določenih oznak segmentov. Na sliki 5.3 zgornji del prikazuje referenčne oznake posameznih segmentov, spodnji pa samodejno pridobljeno strukturo posnetka. V tem primeru je situacija naslednja: imamo dva referenčna govorca, prvega v segmentu od t_1 do t_2, drugega v segmentu od t_4 do t_7. V času od t_2 do t_4 v referenčnem posnetku nimamo govora. V samodejno označenem posnetku pa je situacija drugačna. V tej hipotetični situaciji smo po segmentaciji in rojenju pridobili tri govorce: govorec$_1$ na segmentu $[t_1, t_3]$, govorec$_2$ na segmentu $[t_4, t_5]$ in govorec$_3$ na segmentu $[t_6, t_7]$ ter dva področja brez govora na segmentih $[t_3, t_4]$ in $[t_5, t_6]$. V primeru, da z govorcem$_1$ opisujemo prvega referenčnega govorca in z govorcem$_2$ drugega, tako pridobimo vse tri napake. Napaka napačnega ujemanja (govorec$_3$) je narejena na intervalu $[t_6, t_7]$, napaka zaznavanja na intervalu $[t_2, t_3]$ je posledica napačne določitve segmenta $[t_2, t_3]$ za govor, ravno nasprotno pa je napaka zgrešitve na intervalu $[t_5, t_6]$ posledica napačne določitve segmenta za področje brez govora.

Pri tem je potrebno predhodno določiti, katere referenčne in samodejne oznake štejemo za pravilne in katere oznake postanejo zato nepravilne. To dosežemo tako, da poiščemo takšno preslikavo med referenčnimi in samodejnimi oznakami, s katero maksimiziramo skupni čas ujemanja referenčnih oznak s samodejnimi. Tako definirana preslikava je injektivna, torej gre za ujemanje ene referenčne oznake samo z eno samodejno oznako. Vse ostale oznake, ki se ne ujemajo, štejemo za napake nepravilnega ujemanja. Določanje preslikave ujemanja običajno izvajamo s postopki požrešnega iskanja (metoda razveji in omeji, [Kozak-97]) in predstavlja računsko najbolj zahteven del pri vrednotenju z mero DER.

Ko pa enkrat določimo preslikavo ujemanja med referenčnimi in samodejnimi oznakami, lahko definiramo *mero DER* kot:

$$DER = \frac{\sum_{s=1}^{S} T(s) \cdot (\max(N_{ref}(s), N_{clust}(s)) - N_{ujemanje}(s))}{\sum_{s=1}^{S} T(s) \cdot N_{ref}(s)}, \tag{5.9}$$

kjer je S število segmentov v posnetku, $T(s)$ čas trajanja segmenta s, $N_{ref}(s)$ število referenčnih govorcev prisotnih v segmentu s, $N_{clust}(s)$ število samodejno določenih govorcev prisotnih v segmentu s, ki smo jih pridobili s postopkom rojenja, in $N_{ujemanje}(s)$ število referenčnih in samodejno določenih govorcev, ki se ujemajo v segmentu s. Tako števec v izrazu (5.9) predstavlja skupno vsoto trajanj vseh treh napak, imenovalec pa skupno trajanje govora posameznih govorcev. Mero DER iz enačbe (5.9) najlažje izračunamo tako, da posnetek, na katerem ocenjujemo ujemanje segmentov, razdelimo na toliko odsekov, kolikor je skupnih (referenčnih + samodejnih) mej v posnetku. Če računamo mero DER na takšnih segmentih, potem je izraz $(\max(N_{ref}(s), N_{clust}(s)) - N_{ujemanje}(s))$ v enačbi (5.9) lahko samo 0 v primeru popolnega ujemanja ali 1 v primeru napake. Pomnoženo s časom trajanja $T(s)$ pa dobimo z vsoto skupno trajanje vseh napak skupaj. Če to delimo s skupnim trajanjem govora, ki ga imamo v števcu izraza (5.9), dobimo delež trajanja napačnih razvrstitev segmentov glede na skupno trajanje govora v posnetku. Manjši kot je delež, boljše je razvrščanje segmentov. Delež pa lahko tudi preseže vrednost 1.0, če detektiramo preveč ne–govornih delov v primerjavi z govorom v posnetku.

Mero DER skupaj s postopkom iskanja preslikave ujemanja med referenčnimi in samo-

dejnimi oznakami imenujemo s skupnim imenom *metoda DER*.

Za izračun mere DER smo v naših preizkusih uporabljali orodje, ki je bilo razvito za evaluacije v okviru projekta *NIST Rich Transcriptions*[1].

5.5.1.1 Vrednotenje postopkov rojenja na celotnem intervalu rojenja

Za vsak posnetek lahko z metodo DER ocenimo skupno napako razvrščanja segmentov le, če poznamo končno število rojev razvrščanja. Zaradi tega moramo v postopke rojenja vključiti tudi kriterije zaustavitve rojenja. Z drugimi besedami, če želimo primerjati med seboj postopke rojenja z različnimi principi delovanja, moramo to vedno početi v kombinaciji s kriteriji zaustavitve rojenja, kar pa ni nujno povezano. Tako lahko vrednotimo samo celotne sisteme ne pa posamezne komponente oziroma dele postopkov.

V našem primeru smo tako, da bi odpravili to pomanjkljivost, razvili vrednotenje postopkov rojenja na podlagi mere DER na celotnem intervalu rojenja. Ideja je bila v tem, da smo z mero DER ocenjevali napake postopkov rojenja na vsakem koraku združevanja rojev posameznih zvočnih posnetkov. Tako smo na vsakem zvočnem posnetku lahko primerjali posamezne postopke na celotnem intervalu rojenja, torej od prvega koraka rojenja, ko je rojev toliko, kot je segmentov v posnetku, do zadnjega, ko imamo samo en roj v posnetku. Mera DER iz enačbe (5.9) je bila tako odvisna tudi od števila rojev S v posameznem posnetku. S tem smo dosegli dvoje. Lahko smo analizirali delovanje posameznih postopkov rojenja v odvisnosti od števila rojev in primerjali delovanje postopkov med seboj ne glede na kriterije zaustavitve rojenj.

Skupno oceno mere DER na vseh posnetkih pa smo izvedli s povprečenjem mere DER po posameznih posnetkih v okolici števila referenčnih govorcev posameznih posnetkov.

5.5.2 Izvedba preizkusov postopkov rojenja

Kot smo že omenili, smo izvajali preizkuse s tremi postopki rojenja na dveh zbirkah zvočnih posnetkov informativnih oddaj v primeru ročno označenih segmentov in v primeru samodejne segmentacije.

V nadaljevanju bomo opisali izvedbe in posebnosti posameznih postopkov:

Referenčni postopek rojenja (clust_REFBIC): Postopek rojenja je bil izveden na podlagi sheme iz 5.2. Kriterij združevanja je bil kriterij BIC, *bic*, iz enačbe (5.2). λ iz kriterija BIC smo določili na podlagi optimalnih rezultatov razvrščanja na razvojni zbirki. Segmenti so bili predstavljeni z značilkami MFCC in ΔMFCC izračunanih na vsakih 10 ms.

Postopek rojenja MAP–BIC (clust_UBM_MAP): Postopek je bil izveden na podlagi UBM modelov ocenjenih iz zvočnih posnetkov, na katerih smo izvajali

[1] Orodje lahko najdemo na http://www.nist.gov/speech/tests/rt/rt2003/spring/tools/SpkrSegEval-v21.pl

rojenje. Uporabljali smo dva modela UBM: ženski in moški model. Pri tem smo izvedli predhodno razdelitev segmentov po spolu s postopkom razvrščanja z GMM modeli (GMM modele razvrščanja smo naučili na razvojnih posnetkih.). Učili smo 16 mešanic Gaussovih porazdelitev z diagonalnimi kovariančnimi matrikami. Segmenti so bili predstavljeni z značilkami MFCC in ΔMFCC. Na vsakem segmentu smo izvedli normalizacijo značilk s postopki CMVN in FW. Ker smo dobili boljše rezultate razvrščanja z uporabo postopka FW, bomo v nadaljevanju predstavili rezultate samo ob uporabi normalizacije FW. Postopek MAP adaptacije UBM modelov v GMM modele vsakega segmenta je bil izveden v dveh korakih s popravljanjem samo povprečnih vrednosti Gaussovih porazdelitev GMM modelov. Testirali smo dva kriterija združevanja: mero CLR iz enačbe (5.3) in kriterij BIC iz enačbe (5.5). S kriterijem *upkl2* iz enačbe (5.4) smo dobili boljše rezultate kot s kriterijem CLR in slabše kot s kriterijem BIC, zato bodo v nadaljevanju predstavljeni samo rezultati s kriterijem BIC.

Postopek rojenja s fuzijo (clust_FUSION): Postopek rojenja s fuzijo akustičnih in prozodičnih predstavitev segmentov je bil izveden na podlagi postopka opisanega v razdelku 5.4. Akustične značilke so bile MFCC in ΔMFCC izračunane na vsakih 10 ms, prozodične predstavitve so bile izvedene na 10-ih osnovnih prozodičnih značilkah s postopkom opisanem v razdelku 5.4.3. Kriterij združevanja segmentov s fuzijo je bil fuz iz enačbe (5.8). λ iz kriterija BIC, število lastnih vektorjev iz kriterija združevanja segmentov na podlagi prozodičnih značilk in faktor uteži fw so bili izbrani na podlagi optimalnih rezultatov razvrščanja iz razvojne zbirke.

5.5.2.1 Izbira optimalnih parametrov postopkov rojenja na razvojni zbirki

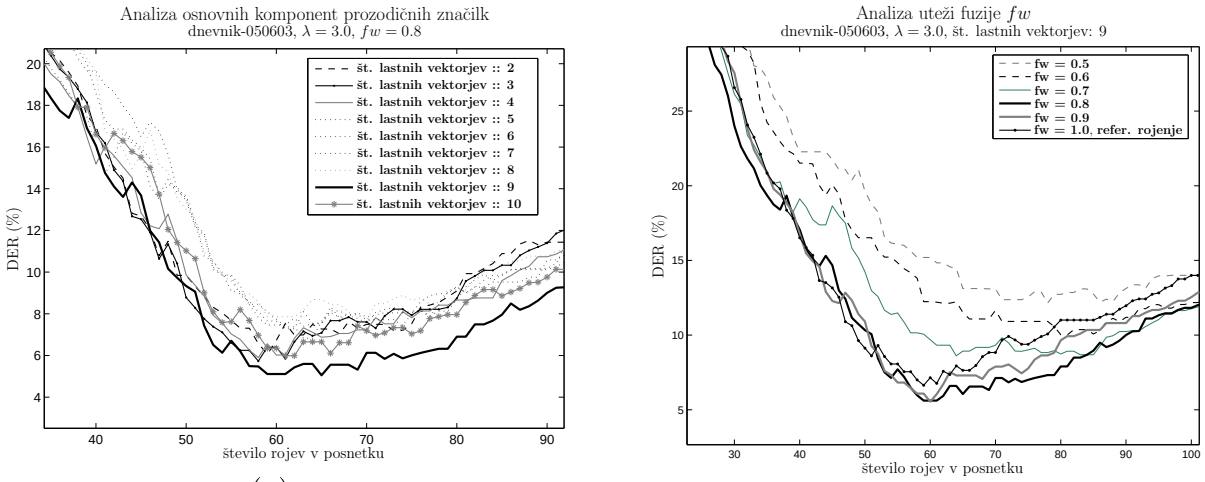
Razvojna zbirka[2], ki smo jo uporabili predvsem za določitev parametrov postopka rojenja s fuzijo, je bila enaka kot v primeru določanja parametrov postopkov segmentacije. Tu smo izvajali postopke rojenja na ročno označenih segmentih zvočnih posnetkov, torej v primeru idealne segmentacije.

Pri referenčnem postopku *clust_REFBIC* smo določali λ iz kriterija BIC. Na podlagi najboljših rezultatov smo izbrali $\lambda = 3.0$.

V primeru postopka *clust_UBM_MAP* smo se na razvojni zbirki odločali o tipu normalizacije, o izbiri kriterija združevanja in o številu Gaussovih porazdelitev v GMM modelih. Na podlagi rezultatov razvrščanja smo se odločili za normalizacijo značilk z metodo FW in postavili število porazdelitev na 16. Med predlaganimi merami združevanja pa smo poleg referenčnega kriterija CLR izbrali še kriterij BIC. Pri tem moramo omeniti, da smo s kriterijem *upkl2* presegli rezultate razvrščanja na vseh posnetkih razvojne zbirke v primerjavi z osnovnim kriterijem CLR, vendar so bili slabši od kriterija BIC ob izbiri $\lambda = 3.0$. Dodatno smo na razvojni zbirki tudi naučili GMM modele ženskega in moškega govora, ki smo jih uporabljali pri razvrščanju segmentov glede na spol za izgradnjo UBM modelov.

[2]Razdelitev zvočnih posnetkov iz zbirk SiBN in COST278 med razvojne in testne posnetke je natančnejše opisana v dodatku A disertacije.

Analiza rezultatov rojenja s fuzijo pri različnih izbirah lastnih vektorjev in uteži fuzije fw.



Slika 5.4: Analiza števila lastnih vektorjev PCA analize prozodičnih značilk (a) in faktorja uteži fw (b) pri rojenju s fuzijo na eni uri zvočnega posnetka *dnevnik-050603* ob idealni segmentaciji.

Pri rojenju s fuzijo smo na podlagi analize rezultatov določili število lastnih vektorjev za izračun kriterija *proz* iz enačbe (5.7) in izbrali primerno utež fw fuzije iz kriterija fuz v enačbi (5.8). λ za kriterij združevanja akustičnih značilk je bil enak kot pri osnovnem referenčnem rojenju, torej $\lambda = 3.0$.

V prvi fazi postopka rojenja s fuzijo smo določali število lastnih vektorjev osnovnih komponent prozodičnih značilk. Na sliki 5.4 (a) je prikazan potek napake razvrščanja, merjen z metodo DER (nižji kot je DER, boljše rojenje pričakujemo) na eni uri posnetka iz razvojne zbirke ob izbiri različnega števila lastnih vektorjev. Pri tem je bila utež fuzije $fw = 0.8$. Na prikazanem posnetku smo najboljše rezultate rojenja dobili v primeru devetih lastnih vektorjev. Rezultati DER so bili ob izbiri devetih lastnih vektorjev v povprečju od 0.5 do 1.5% nižji kot v primeru druge najboljše predstavitve s tremi vektorji. Podobno je bilo tudi na ostalih posnetkih iz razvojne zbirke, zato smo za izračun kriterija *proz* uporabljali 9 lastnih vektorjev iz PCA analize vsakega posnetka.

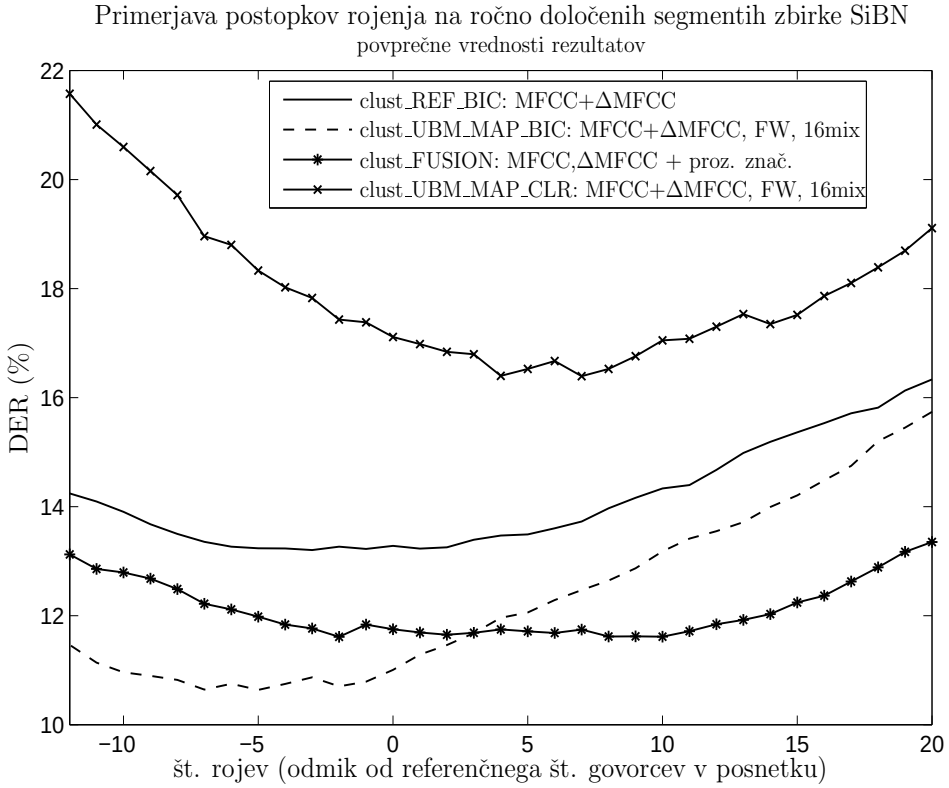
V drugi fazi postopka rojenja s fuzijo smo v primeru devetih lastnih vektorjev izbirali optimalni faktor uteži fw. Vrednosti fw smo izbirali na intervalu od 0.5 do 1.0. V primeru manjših uteži od 0.5 so bili rezultati rojenja slabši in niso prikazani na sliki 5.4 (b), ki tako prikazuje potek rezultatov razvrščanja na eni uri posnetka samo ob izbiri uteži fw iz intervala $[0.5, 1.0]$ s korakom izbire 0.1. V primeru, ko je $fw = 1.0$, imamo rojenje samo na podlagi akustičnih predstavitev, torej je enako rojenju kot pri osnovnem postopku. Na sliki 5.4 (b) lahko opazimo, da smo na prikazanem posnetku presegli rezultate razvrščanja osnovnega postopka v primeru, ko je bil $fw = 0.8$ in $fw = 0.9$. V primeru $fw = 0.8$ smo dosegli najboljše rezultate na tem posnetku. Na ostalih posnetkih iz razvojne zbirke so se najboljši rezultati razvrščanja izmenjavali pri izbirah 0.8 in 0.9, zato smo za testne eksperimente določili $fw = 0.85$.

5.5.3 Primerjava postopkov rojenja v primeru idealne segmentacije

V prvi skupini preizkusov smo primerjali postopke *clust_REF_BIC*, *clust_UBM_MAP* in *clust_FUSION* v primeru idealne segmentacije na testnih posnetkih iz zbirke SiBN. To pomeni, da smo pri razvrščanju segmentov rojili samo tiste odseke zvočnih posnetkov, ki so bili označeni kot govor in pri katerih smo ročno določili meje segmentacije. Tako smo želeli primerjati delovanje postopkov rojenja v idealni situaciji, ko lahko predpostavimo, da nimamo napak predhodne obdelave zvočnih posnetkov.

Testna zbirka SiBN je bila enaka kot v primeru preizkusov segmentacije iz prejšnjega poglavja.

V primeru rojenja s postopkom *clust_UBM_MAP* smo testirali kriterija CLR in BIC, zato smo postopka označevali s *clust_UBM_MAP_CLR* oziroma s *clust_UBM_MAP_BIC*.



Slika 5.5: Primerjava postopkov rojenja na ročno označenih segmentih zbirke SiBN.

Na sliki 5.5 so prikazani skupni rezultati vseh postopkov rojenja. Rezultati na sliki predstavljajo povprečne vrednosti mere DER iz enačbe (5.9) po vseh testnih posnetkih iz zbirke SiBN. Kot smo že omenili, smo vrednosti DER računali na celotnem intervalu rojenja v vsakem posnetku. Povprečne vrednosti DER smo dobili tako, da smo posnetke poravnali glede na dejansko število govorcev v vsakem posnetku in povprečili vrednosti DER glede na enake odmike od dejanskega števila govorcev. Tako vrednost 0 pomeni dejansko število govorcev, vrednost -1 en roj manj, kot je dejansko število govorcev, vrednost 1 en roj več itn.

Kot lahko vidimo na sliki 5.5, so v splošnem rezultati razvrščanja boljši s postopki rojenja *clust_UBM_MAP_BIC* in *clust_FUSION* v primerjavi z referenčnim postopkom *clust_REF_BIC* in postopkom *clust_UBM_MAP_CLR*. Ugotovimo lahko še, da smo s postopkom *clust_UBM_MAP_BIC* v povprečju dobili najboljše rezultate v primeru, ko imamo rojev manj, kot je dejanskih govorcev, in s postopkom *clust_FUSION* najboljše rezultate v primeru, ko je rojev več. Rezultati obeh postopkov so v splošnem boljši od rezultatov referenčnega postopka *clust_REF_BIC*. Če primerjamo sorodne postopke med seboj, lahko ugotovimo, da smo z dodajanjem prozodične informacije v postopku *clust_FUSION* uspeli izboljšati razvrščanje segmentov po govorcih v primerjavi samo z uporabo akustične informacije (postopek *clust_REF_BIC*). Poleg tega je na sliki 5.5 lepo razvidna tudi razlika med postopkoma *clust_UBM_MAP_CLR* in *clust_UBM_MAP_BIC*. Razlika v rezultatih razvrščanja med obema postopkoma se na celotnem intervalu rojenja giblje med 7% in 12%. Tako lahko sklepamo, da z uporabo kriterija BIC za združevanje rojev v postopku *clust_UBM_MAP* izrazito izboljšamo razvrščanje segmentov v primerjavi z osnovnim kriterijem CLR.

Po drugi strani lahko primerjamo delovanje postopkov tudi na podlagi analize najboljših rezultatov posameznih postopkov. Tako lahko ugotovimo, da najboljše rezultate (z mero DER) dobimo v širši okolici dejanskega števila govorcev vsakega posnetka (točka 0). V primeru postopkov *clust_UBM_MAP_BIC* in *clust_REF_BIC* dosežemo minimum v okolici točke -5, medtem ko je minimum v primerih *clust_UBM_MAP_CLR* in *clust_FUSION* v okolici točke $+5$. To pomeni, da v primeru postopka *clust_REF_BIC*, predvsem pa s postopkom *clust_UBM_MAP_BIC* težimo bolj k združevanju rojev na podlagi akustičnih podobnosti (podobne akustične razmere med roji), kar je razumljivo, saj v obeh primerih uporabljamo samo akustične predstavitve segmentov. Medtem ko v primeru postopka *clust_FUSION* z dodano prozodično informacijo poteka rojenje bolj na podlagi govorčevih lastnosti. Napake v postopku so verjetno posledica krajših začetnih segmentov in s tem slabih ocen parametrov prozodičnih značilk.
Druga lastnost delovanja postopkov je razvidna iz poteka vrednosti DER. Tako lahko ugotovimo, da imamo v primeru postopka *clust_FUSION* najmanj strmo padanje vrednosti DER k minimumu v primerjavi z drugimi postopki. To nam zagotavlja sorazmerno dobre rezultate razvrščanja tudi v primeru, ko ne moremo dovolj natančno oceniti optimalnega števila rojev zaustavitve rojenja. Pri vseh ostalih postopkih lahko zaznamo večje padanje vrednosti DER, kar posledično pomeni slabe rezultate razvrščanja v primeru neoptimalnih izbir števila rojev zaustavitve rojenja.

V primeru idealne segmentacije je potrebno omeniti še, da k vrednostim DER prispevajo le napake ujemanja med govorci, saj imamo ne–govorne dele v referenčnem in testnem primeru enake. Ravno zaradi tega so v splošnem rezultati razvrščanja zelo dobri v primerjavi z rezultati iz podobnih eksperimentov [Ajmera-04].

V zaključku lahko podamo sklepno ugotovitev, da smo z izboljšavami obstoječih postopkov rojenja in z vpeljavo dodatne informacije v postopke uspeli izboljšati rezultate razvrščanja segmentov v primeru idealne segmentacije.

5.5.4 Primerjava postopkov rojenja v primeru samodejne segmentacije

V primeru preizkusov rojenja s samodejno segmentacijo je potekalo razvrščanje segmentov v več fazah. V prvi fazi smo pripravili segmente zvočnih posnetkov, v drugi pa smo primerjali rojenje teh segmentov z različnimi postopki rojenja. Za segmentacijo zvočnih posnetkov v prvi fazi smo uporabili postopek segmentacije s fuzijo, s katerim smo v preizkusih iz prejšnjega poglavja dosegali najboljše rezultate segmentacije po govorcih. V tej fazi smo po segmentaciji izvedli še klasifikacijo segmentov na govorne in ne–govorne dele. To smo izvedli s postopkom GNG segmentacije *BICseg–GMM* na podlagi CVS značilk, ki smo jih opisali v poglavju 3. Na tako pripravljenih zvočnih posnetkih smo potem izvajali rojenje s tremi postopki podobno kot v primeru idealne segmentacije.

V tem primeru tako z mero DER merimo poleg napak ujemanja rojev z govorci tudi napake v segmentaciji ter napake v detekciji govornih in ne–govornih delov.

Preizkuse razvrščanja smo izvajali na testnem delu zbirke SiBN in na večjezični zbirki COST278[3].

5.5.4.1 Primerjava postopkov rojenja na zbirki SiBN

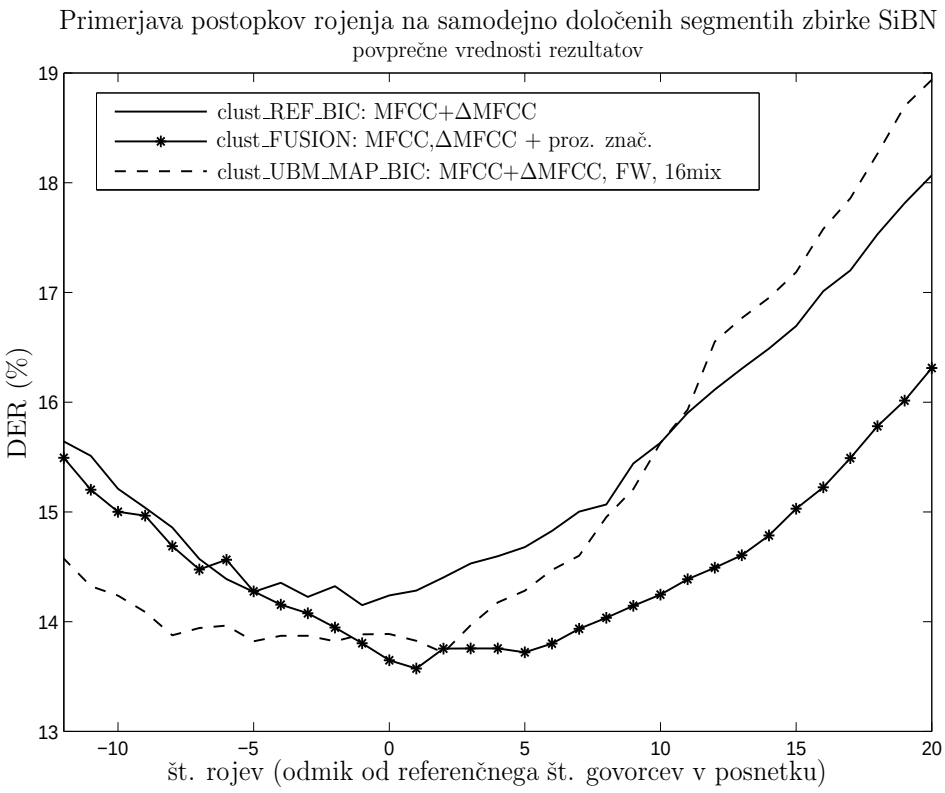
Zbirka SiBN, ki smo jo uporabili za primerjavo postopkov razvrščanja, je bila enaka kot v primeru preizkusov segmentacije v prejšnjem poglavju.

Rezultati postopkov rojenja *clust_REF_BIC*, *clust_UBM_MAP_BIC* in *clust_FUSION* so prikazani na sliki 5.6. Postopka *clust_UBM_MAP_CLR* nismo izvajali, ker smo z njim že v primeru idealne segmentacije dosegli slabše rezultate od ostalih treh postopkov.

Tudi iz rezultatov na sliki 5.6 lahko sklepamo podobno kot v primeru idealne segmentacije. Postopka *clust_UBM_MAP_BIC* in *clust_FUSION* delujeta v povprečju boljše od referenčnega postopka *clust_REF_BIC*. Tudi tu je lepo razvidno, da dobimo najmanjše napake razvrščanja segmentov v okolici dejanskega števila govorcev v posameznem posnetku. V okolici točke 0 smo s postopkom *clust_FUSION* dosegli tudi absolutno najmanjšo napako razvrščanja, ki pa je približno za 3% slabša od najboljših rezultatov v primeru idealne segmentacije.

Potek napak v vseh postopkih je podoben kot v primeru postopkov ob idealni segmentaciji. V primeru postopka rojenja s fuzijo (*clust_FUSION*) je lepo razvidno izboljšanje razvrščanja govornih odsekov v primerjavi z osnovnim postopkom razvrščanja *clust_REF_BIC*. Iz tega lahko sklepamo, da smo z dodajanjem prozodične informacije bistveno izboljšali rojenje po govorcih.
Delovanje sistema *clust_UBM_MAP_BIC* pa je podobno kot v primeru idealne segmentacije, le da je padanje rezultatov DER v tem primeru bolj izrazito kot v primeru

[3]Razdelitev posnetkov iz zbirk SiBN in COST278 med razvojne in testne posnetke je bila enaka kot pri segmentaciji in je natančnejše opisana v dodatku A disertacije.



Slika 5.6: Primerjava postopkov rojenja na samodejno pridobljenih segmentih zbirke SiBN.

idealne segmentacije.

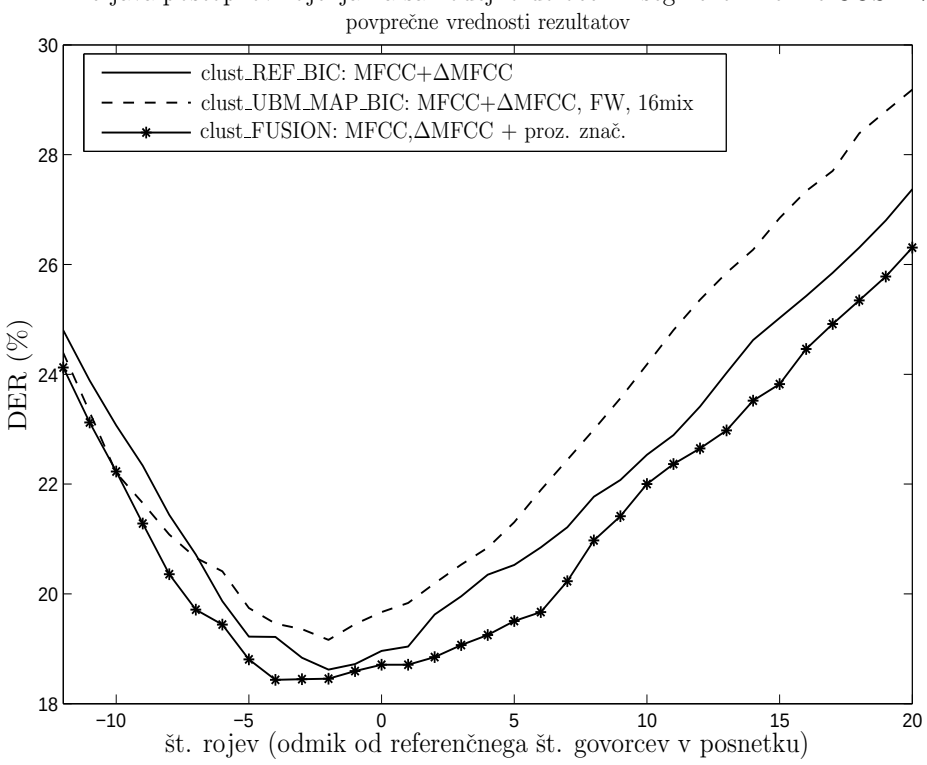
V povprečju so rezultati razvrščanja za 3 ali več odstotkov slabši v primerjavi z rezultati postopkov ob idealnih segmentacijah. To lahko pripišemo predvsem napakam v segmentaciji ter detekciji govornih in ne–govornih delov. To pomeni, da lahko s temi rezultati dodatno ocenjujemo tudi delovanje postopka segmentacije, ki smo ga uporabili v vseh treh primerih rojenja. Glede na to, da so razmerja med postopki v poteku rezultatov ostala nespremenjena v primerjavi z idealno segmentacijo, lahko zaključimo, da z uporabljenim postopkom segmentacije zadovoljivo določamo meje govornih odsekov po govorcih.

5.5.4.2 Primerjava postopkov rojenja na zbirki COST278

Tudi zbirka COST278 je bila enaka zbirki, ki smo jo uporabljali za primerjavo postopkov segmentacije v prejšnjem poglavju.

Na sliki 5.7 so prikazani rezultati rojenja z vsemi tremi postopki.

Tu je še bolj nazorno prikazano dobro delovanje postopka *clust_FUSION* v primerjavi z osnovnim postopkom *clust_REF_BIC*. Na sliki 5.7 lahko opazimo, da je delovanje postopka rojenja s fuzijo boljše na celotnem intervalu rojenja, prav tako pa presega tudi delovanje postopka *clust_UBM_MAP_BIC*, ki je bilo v tem primeru slabše od referenčnega postopka.



Slika 5.7: Primerjava postopkov rojenja na samodejno pridobljenih segmentih zbirke COST278.

Razloge za takšne rezultate lahko pripišemo dejstvu, da je zbirka COST278 bolj akustično raznolika od zbirke SiBN. To pomeni, da je potrebno v primeru rojenja po govorcih združevati roje istih govorcev, ki so akustično manj podobni. Zaradi tega je združevanje govornih odsekov samo na podlagi akustične informacije (*clust_REF_BIC*) oziroma z natančnim modeliranjem akustičnih predstavitev (*clust_UBM_MAP_BIC*) manj zanesljivo kot združevanje z dodano prozodično informacijo.

Tako lahko zaključimo, da se je dodajanje prozodične informacije izkazalo za uspešno pri razvrščanju segmentov v različnih pogojih delovanja in pri različni kvaliteti zvočnih posnetkov. Postopek rojenja s fuzijo je bil najboljši v vseh treh primerih preizkusov. To lahko pripišemo dejstvu, da smo z dodajanjem prozodične informacije uspešno preusmerili rojenje z združevanjem segmentov na podlagi akustičnih podobnosti na rojenje z združevanjem segmentov na podlagi govorčevih lastnosti, kar je bil tudi naš cilj.

5.6 Kriteriji zaustavitve rojenja

Postopki rojenja, ki smo jih primerjali, spadajo med hierarhične postopke rojenja z združevanjem. Združevanje rojev poteka od spodaj navzgor, se pravi, da poteka rojenje od maksimalnega števila rojev do samo enega roja, v katerem so združeni vsi segmenti danega zvočnega posnetka. Pri razvrščanju segmentov po govorcih nas se-

veda ne zanima samo, kako dobro deluje postopek rojenja, ampak tudi kdaj se je potrebno v takem postopku zaustaviti, da dosežemo optimalne rezultate razvrščanja.

V prejšnjih razdelkih smo predvsem preučevali delovanje postopkov rojenja pri različnih izbirah predstavitev govornih segmentov in pri različnih izbirah kriterija združevanja rojev. Pri tem smo iskali optimalno kombinacijo predstavitve govornih odsekov in kriterija združevanja za razvrščanje segmentov po govorcih. V nadaljevanju pa se bomo posvetili kriterijem zaustavitve takšnega tipa rojenja. Pri tem bo poudarek na iskanju takšnega kriterija zaustavitve oziroma iskanju optimalnega števila rojev, pri katerem bi dosegli najboljše rezultate razvrščanja glede na postopek rojenja, ki ga uporabljamo.

Problem zaustavitve rojenja v primeru razvrščanja segmentov po govorcih je problem iskanja števila rojev, pri katerem se postopek rojenja konča. Pri tem moramo poudariti, da iščemo takšno končno število rojev, pri katerem bomo dosegli najboljše rezultate razvrščanja. Običajna praksa pri takšnem rojenju je, da z različnimi merami iščemo *naravno število* rojev, tj. takšno število, pri katerem dosežemo optimalno pokritje zgostitev vzorcev v dani množici vzorcev, [Pavešić-00, str. 247], [Fraley-98]. Mer za ocenjevanje naravnega števila rojev je več in so opisane v [Tibshirani-00, Dudoit-02, Gordon-99]. V primeru razvrščanja segmentov po govorcih pa moramo iskati kriterije zaustavitve na podlagi optimalnih rezultatov razvrščanja in ne na podlagi naravnega števila rojev. Izkazalo se je namreč, da z optimalno izbiro števila rojev zaustavitve lahko znatno vplivamo na rezultate razvrščanja znotraj istega postopka rojenja, kar je lepo razvidno tudi iz rezultatov na slikah 5.5, 5.6 in 5.7. Ob neoptimalnih izbirah končnega števila rojev je lahko rezultat iste metode tudi za nekaj 10% slabši od rezultata ob optimalni izbiri končnega števila rojev.

V nadaljevanju si bomo zato pogledali referenčni kriterij zaustavitve, ki temelji na določanju praga odločitve, kdaj prenehamo z rojenjem, in dva alternativna kriterija, pri katerih iščemo končno število rojev glede na kriterij združevanja rojev oziroma glede na mero DER, s katero merimo napake razvrščanja segmentov.

5.6.1 Obstoječi kriteriji zaustavitve rojenja

V našem primeru razvrščanja segmentov po govorcih je bil cilj razvrstiti segmente tako, da z vsako skupino segmentov opišemo natančno enega govorca. Takšne skupine oziroma roje smo iskali s pomočjo različnih postopkov rojenja segmentov. Te smo predhodno pridobili s postopki segmentacije. Tako pri segmentaciji kot pri rojenju smo se držali osnovnega pravila, da smo določali in združevali segmente na podlagi govorčevih lastnosti. Zato je bilo tu osnovno vodilo, da mora biti tudi kriterij zaustavitve takšnega rojenja v največji meri odvisen od števila dejanskih govorcev v zvočnem posnetku, ki ga obdelujemo.

To lahko dosežemo na več načinov. Ena možnost je ta, da govorne odseke, ki jih združujemo, opisujemo s takšnimi predstavitvami, ki bi v največji možni meri razlikovale različne govorce med seboj in hkrati čim bolj združevale segmente istih govorcev. To pa z drugimi besedami pomeni, da bi množico segmentov zvočnega posnetka zgostili v skupine govornih odsekov istih govorcev. V takem primeru bi potem za kriterije

zaustavitve uporabili mere za iskanje naravnega števila rojev, s katerimi bi ob idealnih razmerah poiskali dejansko število govorcev v posnetku. Zato smo tudi mi v prvem poskusu testirali različne kriterije iskanja naravnega števila rojev. Preizkušali smo številne znane mere: kriterij dvovrstnega korelacijskega koeficienta [Pavešić-00, str. 247], indeks CH [Calinski-74], indeks KL [Krzanowski-85], Hartiganovo statistiko [Hartigan-85], mero obrisa (*ang. silhouette statistic*) [Kaufman-99] in mero GAP (*ang. gap statistic*) [Tibshirani-00]. Vse mere delujejo približno na enak način. Pri vsaki meri na podlagi osnovne mere podobnosti (uporabljene pri kriteriju združevanja) na različne načine ocenjujemo razmerje med razdaljami vzorcev znotraj posameznega roja in razdaljami med vzorci iz različnih rojev. Tako dobimo oceno, kako dobro smo opisali zgostitve v dani množici vzorcev. Na podlagi optimalne ocene zgostitve potem določimo končno število rojev. V naših preizkusih smo z vsemi merami dobili slabe ocene števila rojev. To smo tudi pričakovali glede na dejstvo, da v našem primeru ne iščemo naravnega števila govorcev, pač pa optimalne rezultate razvrščanja. Drugi razlog pa je v tem, da se je v naših postopkih na vsakem koraku rojenja spreminjala tudi razdalja med roji glede na segmente, ki smo jih vključevali v roje. To pomeni, da so se na vsakem koraku spreminjale tudi ocene zgostitve, zato bi bilo v našem primeru potrebno za pravilno določitev naravnega števila rojev na nek način normirati vrednosti ocen na vsakem koraku, kar pa v ocenah s temi merami ni predvideno.

Zato se v primeru razvrščanja segmentov po govorcih za osnovni kriterij zaustavitve rojenja uporablja bolj preprosto metodo, ki je vezana na kriterij združevanja. V tem primeru namreč določimo prag združevanja [Siegler-97, Chen-98, Žibert-05]. To pomeni, da v primeru rojenja z združevanjem iz sheme 5.2 združujemo roje samo do tistega koraka, pri katerih mera podobnosti g ne presega praga združevanja oziroma mera različnosti ne pade pod prag združevanja. Prag je seveda potrebno določiti na podlagi razvojne zbirke in je odvisen od načina predstavitve govornih segmentov in od kriterija združevanja.

V primeru osnovnega rojenja s kriterijem BIC bi morali po definiciji prag postaviti na 0.0, vendar se je izkazalo, da je potrebno prag dodatno določati na podlagi podatkov, ki jih obdelujemo. Pri tem moramo omeniti, da z uvedbo praga rojenja v kriteriju BIC lahko prevedemo osnovni postopek rojenja z združevanjem na osnovni postopek segmentacije s kriterijem BIC. Razlika je le v tem, da pri rojenju predstavlja prag združevanja odločitev, od kje naprej so segmenti različni (torej pripadajo različnim govorcem), pri segmentaciji pa s pragom segmentacije določimo, do kje pustimo združena dva segmenta in od kje naprej ju razdružimo (torej postavimo mejo).

Tudi v primeru ostalih kriterijev združevanja oziroma mer podobnosti (različnosti) ravnamo podobno.

5.6.2 Predlagani kriteriji zaustavitve rojenja

V predlaganih postopkih kriterijev zaustavitve rojenja smo iskali optimalno število rojev na drugačen način.

Kot smo že omenili, v primeru razvrščanja segmentov ni smiselno uporabljati kriterijev

zaustavitev rojenja, ki temeljijo na iskanju naravnega števila rojev samo na podlagi predstavitev segmentov. To pa predvsem zaradi dejstva, ker uporabljamo postopke rojenja za razvrščanje in merimo napako razvrščanja segmentov po govorcih, ne iščemo pa *naravnega števila* rojev v posnetku.

Zato je pri načrtovanju kriterijev zaustavitve v našem primeru potrebno upoštevati predvsem način združevanja segmentov in kriterij zaustavitve določiti na podlagi rezultatov razvrščanja. To je na nek način upoštevano tudi v primeru osnovnega kriterija zaustavitve, ki temelji na pragu združevanja. Na podlagi kriterija združevanja se namreč določi prag odločitve za zaustavitev rojenja. Vendar ima ta postopek dve osnovni pomanjkljivosti: potrebna je razvojna zbirka za določitev kriterija in ne zagotavlja nam optimalne izbire rojev v primeru spremenljivih pogojev delovanja postopkov rojenja.

V nadaljevanju bomo predstavili dva postopka določitve optimalnega števila rojev, s katerima smo želeli odpraviti pomanjkljivosti osnovne metode.

5.6.2.1 Skupni kriterij BIC

Ker se v večini primerov razvrščanja segmentov po govorcih za rojenje in segmentacijo uporablja kriterij BIC, smo tudi izvedbo kriterija zaustavitve zasnovali na tem kriteriju.

Če pri rojenju segmentov z združevanjem uporabljamo kriterij BIC za mero podobnosti med roji, se v bistvu odločamo, ali je boljše dva roja opisati z enim skupnim modelom ali ju bolje opišemo z dvema ločenima modeloma. V vsakem koraku rojenja potem na podlagi najmanjše vrednosti kriterija BIC združimo dva roja. To pomeni, da na vsakem koraku združimo tista dva roja, pri katerih najmanj izgubimo, če ju opišemo samo z enim modelom. Če na to pogledamo s stališča celotnega posnetka, vsak segment pripada nekemu roju, ki ga opišemo z enim modelom. Torej imamo v vsakem koraku toliko modelov, kolikor je rojev, kar pomeni, da smo celoten posnetek opisali s skupnim modelom, ki je sestavljen iz modelov trenutnih rojev. Zato se lahko v duhu kriterija BIC na vsakem koraku rojenja vprašamo, ali smo s skupnim modelom iz vsakega koraka dovolj dobro opisali segmente celotnega posnetka. Najboljši skupni model bo tisti, pri katerem bomo s kriterijem BIC dosegli najboljše rezultate. Tako bomo s kriterijem BIC poiskali najboljši skupni model, ki opisuje dano segmentacijo celotnega posnetka. Število modelov (število rojev) pri najboljšem skupnem modelu tako postane kandidat za zaustavitev rojenja.

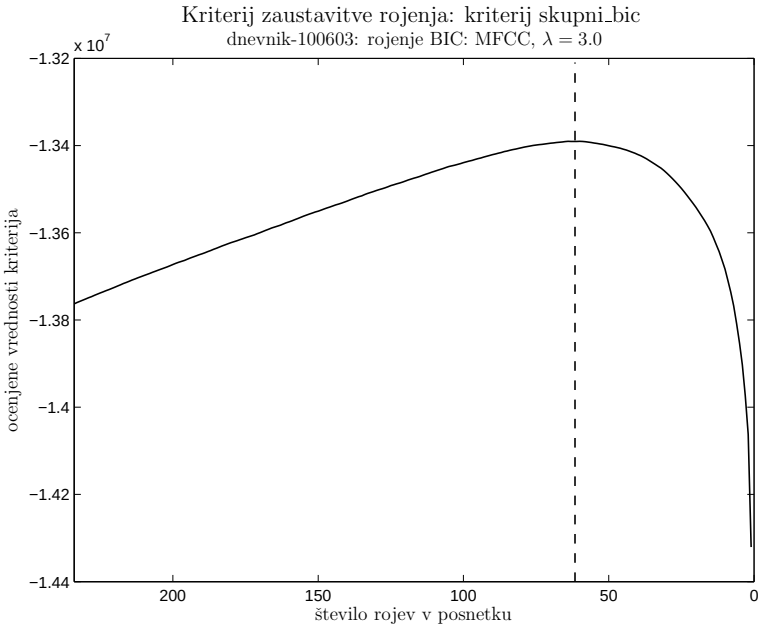
Kriterij BIC, ki smo ga posplošili iz primera izbire enega ali dveh modelov v izbiro med več ali manj modeli, smo poimenovali *skupni kriterij BIC* in ga lahko zapišemo kot:

$$\text{skupni\_bic}(R) = \sum_{r=1}^{R} LLH(M_r) - R \cdot \frac{\lambda}{2} \cdot \#(M_r) \log(N_{vsi}). \tag{5.10}$$

Vsota $\sum_{r=1}^{R} LLH(M_r)$ predstavlja skupno vrednost logaritmov Gaussovih porazdelitev (LLH) v modelih M_r izračunano v točkah predstavitve vseh rojev $r = 1, \ldots, R$ skupaj. Z njo podobno kot pri ostalih verzijah kriterija BIC merimo kvaliteto skupnega modela na danih podatkih (v našem primeru na segmentih celotnega posnetka). Drugi del izraza (5.10) predstavlja kompleksnost modela: R je število vseh modelov (rojev),

$\#(M_r)$ število parametrov modela M_r in N_{vsi} število vseh vzorcev (vektorjev značilk) v posnetku.

Na sliki 5.8 je prikazan primer delovanja skupnega kriterija BIC na posnetku ene informativne oddaje. Število rojev R, pri katerem s kriterijem (5.10) dosežemo maksimalno vrednost, je kandidat za zaustavitev rojenja. Z navpično črtkano premico smo označili dejansko število govorcev v posnetku.



Slika 5.8: Primer delovanja kriterija skupnega BIC na posnetku ene informativne oddaje. Točka maksimalne ocenjene vrednosti kriterija je kandidat za zaustavitev rojenja. Navpična črtkana premica predstavlja dejansko število govorcev v tem posnetku.

Kot je razvidno na sliki 5.8, s kriterijem skupni BIC vedno dosežemo maksimalno vrednost v neki točki rojenja. Naša predpostavka je, da v tej točki dosežemo tudi najbolj optimalne rezultate razvrščanja segmentov po govorcih. To lahko utemeljimo na sledeč način. Če predpostavimo, da govorne odseke, ki pripadajo istemu govorcu, dovolj dobro opišemo z eno Gaussovo porazdelitvijo, torej z enim modelom, potem bi v primeru večjega števila rojev, kot je dejanskih govorcev, dobili bolj natančen skupni model (boljše ocene LLH) segmentov celotnega posnetka, vendar je ta model tudi bolj kompleksen (večje število parametrov). V primeru manjšega števila rojev, pa bi dobili manj natančen skupni model z manjšo kompleksnostjo. V primeru pravilnega rojenja (združevanja govornih odsekov, ki pripadajo istemu govorcu) to pomeni, da dosežemo maksimalno vrednost skupnega kriterija BIC ravno pri številu modelov, ki je enako številu dejanskih govorcev v posnetku. V primeru večjega števila rojev kot je dejanskih govorcev, bi namreč s skupnim kriterijem BIC dobili manjše vrednosti, kot če bi združili dva roja, ki pripadata istemu govorcu (torej ju je bolje opisati z enim samim modelom kot pa z dvema ločenima). V primeru manjšega števila rojev kot je govorcev, pa je ravno obratno. Če združimo dva roja, ki pripadata različnima govorcema, potem je bolje, če roja opišemo z dvema modeloma kot pa s skupnim, kar v primeru skupnega kriterija BIC pomeni, da dobimo večje vrednosti, če imamo ločena

modela, kot pa enega skupnega.

Seveda je skupni kriterij BIC odvisen od pravilnega delovanja kriterija združevanja. V primeru združevanja rojev s kriterijem BIC se oba kriterija BIC zaustavitve in združevanja dopolnjujeta. Združevanje s kriterijem BIC namreč pomeni, da pri rojenju združujemo tiste roje, ki jih je bolje opisati z enim skupnim modelom kot pa z ločenimi. Na vsakem koraku, torej med pari rojev, poiščemo tista roja, pri katerih najmanj izgubimo, če ju opišemo s skupnim modelom. S tem pa hkrati tudi največ pridobimo v primeru skupnega kriterija BIC. Zaustavimo pa se takrat, ko s kriterijem BIC sicer združimo dva najbolj primerna roja, vendar z združitvijo poslabšamo skupni model. Tako lahko gledamo na združevanje s kriterijem BIC kot na odločanje na podlagi lokalnih modelov, s skupnim kriterijem BIC pa se odločamo na podlagi globalnih modelov.

Tako kot pri kriteriju BIC, uporabljenem pri združevanju rojev, je tudi pri skupnem kriteriju BIC potrebno določiti vrednost λ. Tu nam λ ne predstavlja praga odločitve, ampak nam definira položaj maksimalne vrednosti kriterija glede na dano predstavitev segmentov v posnetku. Večja kot je λ, bolj favoriziramo večje število modelov, manjša kot je λ, manjše je število modelov. V splošnem se je izkazalo, da je λ pri skupnem BIC najbolje izbrati enako kot pri združevanju s kriterijem BIC. Sicer pa tudi tu določamo λ na podlagi optimalnih rezultatov razvrščanja segmentov iz razvojnih zbirk.

5.6.2.2 Relativni DER z dvema rojenjema

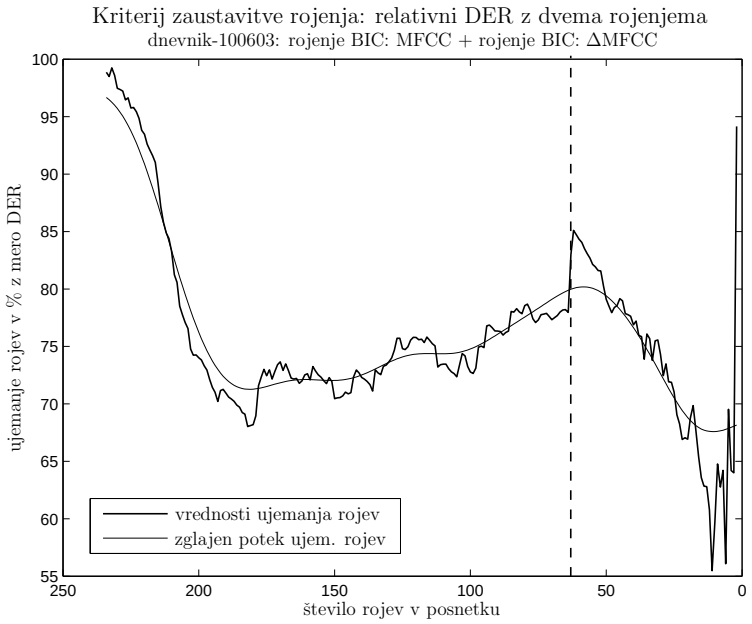
Povsem drugačen pristop pri iskanju optimalnega števila rojev smo uporabili v postopku iskanja z uporabo mere napake DER, opisane v razdelku 5.5.1.1. Tu smo hoteli poiskati optimalno število rojev glede na rezultate razvrščanja, ne pa glede na kriterij združevanja, kot je bilo to v primeru skupnega kriterija BIC.

Osnovno vodilo pri tem je bilo, da bi za določitev optimalnega števila rojev uporabili kar rezultate razvrščanja. To bi v našem primeru pomenilo, da bi izbirali med tistimi števili rojev v posnetku, pri katerih bi z mero DER dosegli najboljše rezultate razvrščanja. Pri tem bi uporabljali primerjavo rezultatov razvrščanja na celotnem intervalu rojenja, kar smo že predstavili v razdelku 5.5.1.1.

Pri vrednotenju postopkov razvrščanja segmentov je seveda potrebno imeti referenčne oznake posnetkov, ki jih primerjamo s samodejno pridobljenimi. V našem primeru tako z mero DER primerjamo oznake, ki smo jih pridobili na podlagi nekega postopka razvrščanja segmentov, z oznakami, ki jih ročno označimo. Ker pa v primeru samodejnega razvrščanja običajno ne razpolagamo z referenčnimi oznakami (saj je to ravno cilj razvrščanja), smo se odločili, da bi z mero DER primerjali dva različna postopka rojenja in bi na ta način poskusili ugotoviti število rojev optimalnega razvrščanja. Ideja je bila v tem, da bi ob enakih pogojih segmentacije in detekcije govora uporabili dva različna (neodvisna) postopka rojenja, ki bi ju z mero DER primerjali na celotnem intervalu rojenja.

Tipičen potek primerjave rezultatov z mero DER je prikazan na sliki 5.9.

Na začetku rojenja z združevanjem segmentov imamo v primeru obeh postopkov roje-



Slika 5.9: Primer delovanja predlaganega kriterija relativnega DER z dvema rojenjema na posnetku ene informativne oddaje. Točka maksimuma med dvema lokalnima minimumoma (na zglajeni verziji kriterijske funkcije) je kandidat za zaustavitev rojenja. Navpična črtkana premica predstavlja dejansko število govorcev v tem posnetku.

nja enake roje. Vsak segment je namreč predstavljen s svojim rojem v obeh primerih, zato je tudi napaka z mero DER 0 oziroma je ujemanje[4] 100%. Podobno je tudi na koncu rojenja, ko imamo v obeh primerih en skupen roj in je zato ujemanje prav tako 100%. Vmes je tipičen potek ujemanja podoben poteku na sliki 5.9, kar lahko razložimo na naslednji način. V primeru, da imamo dva neodvisna postopka rojenja, ki sorazmerno zanesljivo razvrščata segmente po govorcih, potem lahko pričakujemo, da bomo v primeru obeh postopkov dobili najboljše rezultate razvrščanja v okolici dejanskega števila govorcev v posnetku. To pa v našem primeru pomeni, da bo tam tudi največje ujemanje oznak obeh posnetkov. Ob predpostavki, da oba postopka delujeta neodvisno, torej da se na istih korakih rojenja združujejo različni roji iz obeh postopkov, lahko pričakujemo potek ujemanja, kot je prikazan na sliki 5.9. Na začetku je ujemanje 100%; na intervalu od prvega koraka združevanja do dejanskega števila govorcev v posnetku lahko pričakujemo slabše ujemanje, ker se rojenji med seboj razlikujeta; v okolici dejanskega števila posnetkov se rojenji ujemata, zato tam pričakujemo maksimum; v intervalu od dejanskega števila govorcev do enega roja pa lahko pričakujemo spet neujemanje obeh rojenj do zadnjega koraka, ko imamo samo en roj in je ujemanje zato 100%.

Postopek iskanja optimalnega števila rojev na podlagi ujemanja dveh rojenj z mero DER je tako naslednji:

1. izvedemo dva različna postopka rojenja na istih govornih odsekih;

[4]Ujemanje je definirano kot 100%−DER. Ker DER ni omejen med 0 in 100%, je potrebno v tem primeru DER ustrezno normirati.

2. z mero DER primerjamo rezultate enega rojenja z drugim;
3. iščemo lokalni maksimum med dvema minimumoma:
   - funkcijo poteka ujemanja gladimo, dokler ne dobimo samo dveh globalnih minimumov;
   - na zglajeni funkciji poiščemo maksimum na intervalu med dvema minimuma;
4. število rojev, pri katerem zglajena funkcija doseže lokalni maksimum, je predlagano število rojev za zaustavitev rojenja.

V predlaganem postopku iskanja optimalnega števila rojev ne potrebujemo dodatnih razvojnih zbirk za določitev odprtih parametrov kriterija, potrebno pa je izvajanje dveh rojenj in primerjava z mero DER na celotnem intervalu rojenja. Pri tem je bistvenega pomena, da uporabljamo čimbolj različne postopke rojenj, s katerimi na različne načine združevanja segmentov pridemo do skupnega rezultata – razvrstitve segmentov po govorcih. V primeru, da s to metodo ne najdemo ustreznega minimuma vrednosti ujemanja, lahko za določitev končnega števila rojev uporabimo postopek skupnega kriterija BIC, ki smo ga predstavili v prejšnjem razdelku.

5.7 Preizkusi kriterijev zaustavitve rojenja

V nadaljevanju bomo predstavili skupne rezultate rojenja s postopki *clust_REF_BIC*, *clust_UBM_MAP_BIC* in *clust_FUSION* v primeru različnih kriterijev zaustavitve rojenja. Ti rezultati hkrati predstavljajo tudi končne rezultate razvrščanja segmentov zvočnih posnetkov informativnih oddaj po govorcih.

Preizkušali smo tri različne kriterije zaustavitve rojenja: referenčni kriterij, kjer zaustavitev rojenja definiramo s pragom zaustavitve rojenja, *skupni kriterij BIC*, kjer ob izbiri optimalne vrednosti λ iščemo končno število rojev, in postopek iskanja optimalnega števila rojev na podlagi *relativnega DER* z dvema rojenjema. Končne rezultate razvrščanja smo primerjali med seboj in z rezultati, ki smo jih dobili, če smo namesto kriterijev zaustavitve ustavili rojenje kar pri dejanskemu številu govorcev v posnetku.

Optimalne vrednosti zaustavitve rojenj smo iskali na osnovnem postopku rojenja *clust_REF_BIC* ob uporabi značilk MFCC+ΔMFCC. Število rojev, ki smo jih dobili iz osnovnega postopka, smo uporabili kot osnovo za izbiro rojev v ostalih postopkih. V primeru dveh rojenj pri metodi ocenjevanja števila rojev z relativnim DER smo za prvo rojenje uporabili osnovni postopek rojenja *clust_REF_BIC* z uporabo MFCC značilk, drugi postopek pa je bil ravno tako *clust_REF_BIC*, vendar z uporabo samo ΔMFCC značilk.

Končni rezultati razvrščanja segmentov ob uporabi različnih kriterijev zaustavitve rojenja so bili pridobljeni v primeru idealne in v primeru samodejne segmentacije. Slednje lahko štejemo tudi za končne rezultate razvrščanja segmentov s sistemom, kjer so bile vse faze strukturiranja posnetka – segmentacija, GNG razvrščanje in rojenje – izvedene samodejno. Preizkusi primerjave kriterijev zaustavitve rojenja so bili izvedeni na istih zbirkah, ki smo jih uporabili že v prejšnjem primeru primerjave postopkov rojenja.

5.7.1 Primerjava kriterijev zaustavitve v primeru idealne segmentacije na zbirki SiBN

V primeru idealne segmentacije na zbirki SiBN smo primerjali postopke rojenja na celotnem intervalu rojenja na sliki 5.5. Končni rezultati ob uporabi različnih kriterijev zaustavitve pa so predstavljeni v tabeli 5.1.

Tabela 5.1: Končni rezultati rojenja vseh postopkov na ročno označenih segmentih zbirke SiBN glede na optimalne izbire vrednosti kriterijev zaustavitve rojenja. Skupni rezultati so povprečni rezultati DER na vseh posnetkih testne zbirke SiBN.

postopek rojenja	*dejansko št. govorcev* DER (%)	*prag zaustavitve* DER (%)	*skupni BIC* DER (%)	*relativni DER z dvema rojenjema* DER (%)
clust_REF_BIC: MFCC+ΔMFCC	13.28	15.38	13.36	13.18
clust_UBM_MAP_BIC: MFCC+ΔMFCC, FW, UBM: 16mix	11.01	14.29	11.12	10.94
clust_FUSION: MFCC, ΔMFCC + proz. značilke	11.75	12.32	11.57	11.34

Rezultate iz tabele 5.1 lahko beremo na dva načina: po stolpcih in vrsticah. Če primerjamo rezultate po stolpcih, primerjamo rezultate skupne napake razvrščanja glede na posamezne postopke rojenja ob uporabi istega kriterija zaustavitve. Če pa primerjamo rezultate po vrsticah, primerjamo delovanje različnih kriterijev zaustavitve ob uporabi istega postopka rojenja.

V prvem stolpcu so predstavljeni povprečni rezultati DER v primeru, če smo v vsakem posnetku za zaustavitev rojenja uporabili število dejanskih govorcev. Tu lahko ugotovimo, da smo s postopkom *clust_UBM_MAP_BIC* dobili najboljše rezultate, kar je lepo razvidno tudi na sliki 5.5. V drugem stolpcu so zbrani rezultati ob uporabi osnovnega kriterija zaustavitve rojenja z uporabo praga odločitve. Rezultati so primerjalno slabši kot v primeru dejanskega števila govorcev, kar govori v prid dejstvu, da s tem kriterijem ne najdemo najbolj optimalnih točk zaustavitve rojenj v vseh postopkih. Lahko pa tudi opazimo, da je v tem primeru najboljši postopek *clust_FUSION*, iz česar lahko na podlagi primerjave iz slike 5.5 sklepamo, da s pragom zaustavitve v povprečju poiščemo večje število rojev, kot je dejanskih govorcev. Tretji stolpec predstavlja rezultate kriterija skupnega BIC. Rezultati so boljši kot v primeru praga zaustavitve in povsem primerljivi z rezultati v primeru dejanskega števila govorcev. To pomeni, da se s skupnim kriterijem BIC v povprečju povsem približamo ocenam dejanskega števila govorcev v posnetkih. V zadnjem stolpcu so zbrani rezultati razvrščanja ob uporabi postopka relativnega DER na dveh rojenjih. Ti rezultati so najboljši med vsemi, tudi od rezultatov, kjer zaustavitev določamo z dejanskim številom govorcev (prvi stolpec). Razlog za to je predvsem v tem, da tu določamo število rojev glede na optimalne re-

zultate DER in ne na dejansko število govorcev, kot je to v prvem primeru.
Drugače pa razmerja med rezultati vseh treh postopkov v primeru kriterija skupnega BIC in relativnega DER ostajajo nespremenjena.

Tako lahko zaključimo, da s predlaganima postopkoma iskanja končnega števila rojev izboljšamo skupne rezultate razvrščanja vseh postopkov rojenja. V primeru skupnega kriterija BIC, ki je bil načrtovan glede na kriterij združevanja, se povsem približamo rezultatom dejanskega števila govorcev v posnetkih. V primeru postopka relativnega DER z dvema rojenjema, ki temelji na iskanju končnega števila rojev glede na mero napake DER, pa v primeru idealne segmentacije celo izboljšamo skupne rezultate razvrščanja.

5.7.2 Primerjava kriterijev zaustavitve na zbirki SiBN

V nadaljevanju bomo primerjali še končne rezultate postopkov v primeru samodejne segmentacije na zbirkah SiBN in COST278.

Skupni rezultati razvrščanja segmentov na zbirki SiBN so prikazani v tabeli 5.2. Tu so zbrani končni rezultati razvrščanja ob uporabi različnih kriterijev zaustavitve vseh treh postopkov rojenja, ki smo jih primerjali že na sliki 5.6.

Tabela 5.2: Končni rezultati rojenja vseh postopkov na samodejno pridobljenih segmentih zbirke SiBN glede na optimalne izbire vrednosti kriterijev zaustavitve rojenja. Skupni rezultati so povprečni rezultati DER na vseh posnetkih testne zbirke SiBN.

postopek rojenja	*dejansko št. govorcev* DER (%)	*prag zaustavitve* DER (%)	*skupni BIC* DER (%)	*relativni DER z dvema rojenjema* DER (%)
clust_REF_BIC: MFCC+ΔMFCC	14.31	17.01	14.18	14.96
clust_UBM_MAP_BIC: MFCC+ΔMFCC, FW, UBM: 16mix	13.88	17.38	13.49	14.25
clust_FUSION: MFCC, ΔMFCC + proz. značilke	13.59	15.32	13.80	13.79

Tudi iz končnih rezultatov v tabeli 5.2 lahko ugotovimo, da se razmerja med rezultati iz različnih postopkov rojenja ohranjajo, s tem da smo s postopkom *clust_FUSION* dobili v večini primerov najboljše rezultate. To je predvsem posledica delovanja tega postopka na celotnem intervalu rojenja. Kot smo že omenili pri primerjavi delovanja postopkov na sliki 5.6, v primeru rojenja s *clust_FUSION* opazimo najmanj strmo padanje vrednosti DER k minimumu v primerjavi z drugimi postopki. Iz tega sledi, da tudi če s kriterijem zaustavitve zgrešimo minimalno vrednost napake DER, v tem primeru še vedno ostanemo v območju dobrega delovanja postopka. To pa ne velja za postopka

clust_REF_BIC in *clust_UBM_MAP_BIC*, kjer imamo v obeh primerih bolj strmo padanje k minimalni vrednosti napake. Posledica tega je, če s kriterijem zaustavitve ne najdemo optimalnih vrednosti zaustavitve rojenja, je napaka končnih rezultatov večja. To je lepo razvidno v primeru osnovnega kriterija, definiranega s pragom zaustavitve, in v primeru relativnega DER (zadnji stolpec), kjer pa so rezultati razvrščanja boljši. S skupnim kriterijem BIC smo v povprečju zelo dobro ocenili optimalno število rojev za zaustavitev, saj smo v tem primeru celo presegli referenčne rezultate, kjer je bila zaustavitev, definirana z dejanskim številom govorcev. Obenem lahko ugotovimo, da s kriterijem relativnega DER nismo presegli referenčnih rezultatov, kot je bilo to v primeru idealne segmentacije, kar priča o tem, da je ta postopek bolj učinkovit v primeru bolj natančnih segmentacij.

Kljub temu smo z obema predlaganima postopkoma zaustavitve znatno presegli končne rezultate razvrščanja segmentov v primerjavi z osnovnim postopkom, definiranim s pragom zaustavitve.

5.7.3 Primerjava kriterijev zaustavitve na zbirki COST278

Delovanje postopkov rojenja ob različnih kriterijih zaustavitve smo preverjali tudi v primeru zbirke COST278. Na sliki 5.7 smo primerjali delovanje postopkov na celotnem intervalu rojenja, v tabeli 5.3 pa so zbrani končni rezultati ob uporabi različnih kriterijev zaustavitve.

Tabela 5.3: Končni rezultati rojenja vseh postopkov na samodejno pridobljenih segmentih zbirke COST278 glede na optimalne izbire vrednosti kriterijev zaustavitve rojenja. Skupni rezultati so povprečni rezultati DER na vseh posnetkih zbirke COST278.

postopek rojenja	*dejansko št. govorcev* DER (%)	*prag zaustavitve* DER (%)	*skupni BIC* DER (%)	*relativni DER z dvema rojenjema* DER (%)
clust_REF_BIC: MFCC+ΔMFCC	19.02	22.32	19.43	20.00
clust_UBM_MAP_BIC: MFCC+ΔMFCC, FW, UBM: 16mix	19.48	24.11	20.08	20.53
clust_FUSION: MFCC, ΔMFCC + proz. značilke	18.83	21.87	19.29	19.71

Rezultati iz tabele 5.3 ohranjajo podobna razmerja kot v primeru zbirke SiBN, le da imamo tu večja odstopanja med kriteriji zaustavitve. To lahko razložimo z dejstvom, da smo vse odprte parametre postopkov iskanja optimalnega števila rojev ocenjevali na podlagi razvojne zbirke, kjer so bili izbrani posnetki iz zbirke SiBN. Drugi razlog je tudi v tem, da je zbirka COST278 bolj raznolika, zato smo v primeru samodejne segmentacije delali več napak pri segmentaciji in detekciji govora.

Ta dejstva potrjujejo, da smo se s skupnim kriterijem BIC najbolj približali referenčnim rezultatom na podlagi dejanskega števila govorcev v posnetkih. Druga ugotovitev govori v prid postopku rojenja s fuzijo, *clust_FUSION*. Če primerjamo končne rezultate tega postopka ob uporabi različnih kriterijev zaustavitve, lahko ugotovimo, da v povprečju s tem postopkom delamo najmanj napak ne glede na optimalno določene vrednosti zaustavitve rojenja. To si lahko podobno kot v prejšnjih preizkusih razlagamo z dejstvom, da postopek *clust_FUSION* deluje dobro v celotni okolici dejanskega števila govorcev v posameznem posnetku, kar zagotavlja večjo neobčutljivost postopka na neoptimalno določene vrednosti zaustavitve rojenja. Ob tem smo opazili tudi nekoliko slabše rezultate rojenja pri postopku relativnega DER v primerjavi s skupnim kriterijem BIC, kar lahko razložimo podobno kot v prejšnjem primeru, da je kriterij relativnega DER bolj občutljiv na natančnost segmentacije.

Podobno kot v prejšnjih preizkusih lahko tudi tu zaključimo, da smo se z obema predlaganima postopkoma zelo približali referenčnim rezultatom rojenja, kjer smo zaustavitev definirali z dejanskim številom govorcev v posnetkih. V primerjavi z osnovnim kriterijem zaustavitve pa smo v povprečju končne rezultate znatno presegli.

5.8 Zaključek

V tem poglavju smo predstavili postopke razvrščanja govornih odsekov zvočnih posnetkov po govorcih. Razvrščanje segmentov smo izvajali s postopki rojenja, ki skupaj s postopki segmentacije in razvrščanja segmentov na govorne in ne–govorne dele predstavljajo osnovo za strukturiranje zvočnih posnetkov v različnih sistemih govornih tehnologij. Naš namen je bil izboljšati postopke razvrščanja segmentov po govorcih, ki bi jih lahko uporabili v sistemih za samodejno označevanje zvočnih posnetkov informativnih oddaj za različne namene uporabe, bodisi za iskanje in sledenje govorcev v zvočnih zbirkah, za sledenje in pridobivanje informacij o vsebini informativnih oddaj, ali pa v sistemih za indeksacijo zvočnih posnetkov.

V osnovnem postopku razvrščanja segmentov se uporablja postopek rojenja z združevanjem, kjer za mero podobnosti med roji uporabljamo kriterij BIC. V okviru doktorskega dela smo hoteli izboljšati osnovni postopek rojenja na več načinov: z iskanjem boljših predstavitev govornih odsekov skupaj z različnimi kriteriji združevanja, z vpeljavo dodatne informacije pri združevanju segmentov in z izvedbo novih kriterijev za zaustavitev rojenja.

Pri predstavitvi osnovnih govornih odsekov smo iskali takšne lastnosti segmentov oziroma predstavitve, da bi s postopki rojenja združevali samo tiste segmente, ki pripadajo istim govorcem. Hkrati pa smo želeli čimbolj odpraviti neželene pojave različnih akustičnih pogojev. Zato smo predstavili še en alternativen postopek rojenja segmentov po govorcih na podlagi UBM in GMM modelov, ki se uporabljajo v sistemih za razpoznavanje govorcev. V tem primeru smo izvajali združevanje govornih odsekov na podlagi akustičnih predstavitev, ki smo jih predhodno dodatno normalizirali, da bi se znebili vplivov različnih akustičnih pogojev. Z dodatnimi izboljšavami pri tvorbi UBM modelov in predvsem ob zamenjavi osnovnega kriterija združevanja s kriterijem BIC

smo znatno izboljšali rezultate razvrščanja.

Povsem drugačno pot smo izbrali pri drugem postopku rojenja. Tu smo poleg osnovne akustične informacije dodali še prozodično informacijo za predstavitev segmentov. S tem smo hoteli doseči dvoje. Z dodatno prozodično informacijo smo hoteli preusmeriti združevanje govornih odsekov z združevanja samo na podlagi akustičnih podobnosti na združevanje na podlagi govorčevih lastnosti. Drugi namen pa je bil ta, da smo s tem hoteli izboljšati tudi neobčutljivost postopkov združevanja na spremenljive akustične pogoje v zvočnih posnetkih. Dejstvo je namreč, da je prozodična informacija razmeroma neobčutljiva na različne akustične spremembe. V tem primeru smo osnovni postopek rojenja spremenili tako, da je združevanje potekalo na podlagi dveh predstavitev: osnovne akustične in dodatne prozodične. Zato smo predlagali tudi nov kriterij združevanja, ki je temeljil na uteženi meri podobnosti med obema predstavitvama. Parametre kriterijev združevanja smo dodatno analizirali na razvojni zbirki posnetkov.

Vrednotenje postopkov rojenja smo izvajali z mero DER. Tu smo predstavili nov pristop vrednotenja razvrščanja segmentov po govorcih, ki je neodvisen od kriterija zaustavitve rojenja. To smo dosegli tako, da smo izvajali primerjavo postopkov na celotnem intervalu rojenja, kjer smo najprej poravnali rezultate rojenj glede na dejansko število govorcev v posnetkih in s povprečenjem vrednosti DER dobili potek delovanja postopkov rojenja. Na ta način smo lahko primerjali delovanje postopkov neposredno iz potekov napak razvrščanja segmentov po govorcih.

V zadnjem delu poglavja smo se ukvarjali predvsem s kriteriji zaustavitve postopkov rojenja. Tu smo predlagali dva izvirna postopka iskanja optimalnega števila rojev za zaustavitev rojenja. Prvi postopek je izhajal iz mere podobnosti, ki smo jo uporabljali pri združevanju rojev. Imenovali smo ga skupni kriterij BIC. Osnovno vodilo pri tem je bilo, da smo iskali število rojev za zaustavitev rojenja med tistimi števili modelov, s katerimi bi najbolje po kriteriju BIC opisali segmente v posameznem modelu. Drugi kriterij pa je izhajal iz mere DER. Tu smo hoteli poiskati takšno število končnih rojev v posameznem posnetku, da bi optimizirali rezultate skupne napake razvrščanja. To smo naredili tako, da smo primerjali ujemanje delovanja dveh različnih postopkov rojenj na isti segmentaciji posameznega zvočnega posnetka. Na podlagi poteka ujemanja smo potem dobili kandidate za zaustavitev postopka rojenja. Z obema kriterijema smo izboljšali iskanje točk zaustavitev postopkov rojenja in s tem znatno izboljšali skupne končne rezultate razvrščanja segmentov po govorcih.

Skupaj s kriteriji zaustavitve smo lahko izvedli končno analizo postopkov rojenja za namen razvrščanja segmentov po govorcih. Skupni končni rezultati preizkusov izvedeni na dveh zbirkah SiBN in COST278 ob idealni in samodejni segmentaciji govorijo v prid predlaganih postopkov rojenja. Na podlagi rezultatov smo ugotovili, da je bilo delovanje postopka rojenja s fuzijo akustične in prozodične informacije najbolj zanesljivo in neobčutljivo na različne akustične pogoje v posnetkih. Zelo dobre rezultate smo dobili tudi s postopkom rojenja na podlagi GMM in UBM modelov v primeru kriterija BIC.

Tako lahko na podlagi skupnih rezultatov zaključimo, da smo s predlaganima postopkoma izboljšali rezultate razvrščanja v primerjavi z osnovnim referenčnim postopkom. To pa predvsem zaradi dejstva, ker smo v obeh primerih gradili rojenje na združevanju govornih odsekov na podlagi govorčevih lastnosti, ne pa neposredno na akustičnih

podobnostih segmentov. V primeru rojenja s fuzijo smo to dosegli z dodajanjem prozodične informacije, v primeru rojenja z GMM in UBM modeli pa z normalizacijo osnovnih akustičnih značilk.

6 Sklep

6.1 Pregled uporabljenih pristopov
6.2 Pomen doseženih ciljev
6.3 Smernice za nadaljnje delo

V zaključnem poglavju bomo pregledali različne pristope in povzeli bistvene prispevke in izboljšave postopkov, ki smo jih uporabljali pri reševanju nalog detekcije govora, samodejne segmentacije in razvrščanja segmentov po govorcih. Posebno pozornost bomo posvetili tudi vključevanju predlaganih postopkov v različne sisteme govornih tehnologij. Tu se bomo osredotočili predvsem na zasnovo sistema za samodejno indeksacijo zvočnih posnetkov po govorcih, kjer bomo natančnejše opredelili vlogo in pomen postopkov, s katerimi smo se ukvarjali v disertaciji. Sklepni del doktorske disertacije bomo zaključili s smernicami za nadaljnje delo.

6.1 Pregled uporabljenih pristopov

Osnovni cilj, ki smo si ga zadali pri predlaganih postopkih detekcije govora, samodejne segmentacije in razvrščanja segmentov po govorcih je bil, da bi zmanjšali vpliv številnih odprtih parametrov postopkov na njihovo delovanje v različnih akustičnih pogojih. Na ta način smo želeli povečati zanesljivost delovanja posameznih postopkov, ki bi vodili k izboljšanju skupnih rezultatov strukturiranja zvočnih posnetkov informativnih oddaj glede na prisotnost govorcev v posnetkih.

Zato smo predlagali številne izboljšave osnovnih postopkov na različnih nivojih delovanja. V disertaciji smo se tako ukvarjali z različnimi predstavitvami zvočnih signalov, s katerimi smo želeli boljše modelirati tiste lastnosti v signalih, ki so nas zanimale pri posameznih nalogah obdelave zvočnih posnetkov. Osnovnim akustičnim predstavitvam signalov smo dodajali še drugačne informacije oziroma predstavitve, ki so bile izpeljane posredno iz časovno–frekvenčnih predstavitev signalov. Na ta način smo v postopke vpeljali informacijo višjega reda. V primeru detekcije govora so bile to fonetične značilke, v primeru razvrščanja segmentov po govorcih pa prozodične značilke. Poleg tega, da smo s takšnimi predstavitvami boljše opisovali oziroma poudarjali tiste lastnosti v zvočnih posnetkih, ki so nas zanimale, smo z uvedbo informacije višjega reda dosegli tudi večjo neobčutljivost postopkov na različne akustične razmere v zvočnih posnetkih. Drugi način izboljšanja predstavitev je bilo združevanje različnih predstavitev s postopki fuzije. Tako smo v primeru detekcije govora združevali akustične in fonetične značilke, v primeru segmentacije različne skupine akustičnih značilk, v primeru razvrščanja segmentov pa smo osnovnim akustičnim predstavitvam dodali prozodične značilke. V teh primerih je bilo seveda potrebno ustrezno prilagoditi predlagane postopke obdelave zvočnih posnetkov.

Ostalo raziskovalno delo je bilo namenjeno predvsem izboljšavi osnovnih postopkov z namenom povečanja neobčutljivosti na različne pogoje delovanja.

Tako smo se pri detekciji govora ukvarjali z različnimi postopki segmentacije, s katerimi bi bolj učinkovito izvajali segmentacijo na govorne in ne–govorne dele. Izvedli smo dva postopka segmentacije. V obeh primerih smo za razvrščanje uporabili GMM modele, ki smo jih ocenili vnaprej. Pri prvem postopku smo GMM modele vključili v mrežo HMM modelov, segmentacija pa je potekala po postopku Viterbijevega dekodiranja. V drugem primeru pa se je najprej izvajala segmentacija s kriterijem BIC glede na akustične spremembe v zvočnih posnetkih, nato pa razvrščanje z GMM modeli. Medtem ko je prvi postopek bolj primeren za vključevanje v druge sisteme govornih tehnologij, predvsem v sisteme za razpoznavanje govora, ki delujejo na podoben način, pa je bil drugi načrtovan za uporabo s predlaganimi fonetičnimi značilkami.

Pri segmentaciji zvočnih posnetkov glede na zamenjave govorcev in spremembe akustičnega ozadja (segmentacija SAG) smo poskušali izboljšati standardne postopke, da bi bolje delovali v spremenljivih akustičnih razmerah. Pri tem smo se predvsem ukvarjali z določanjem pragov odločitve za meje med segmenti. Tako smo razvili postopek segmentacije, kjer smo prag odločitve določali relativno glede na vrednosti kriterija BIC, ki smo jih sprotno ocenjevali na zvočnih posnetkih, ki smo jih obdelovali. Predlagani postopek je tako deloval v dveh fazah: v prvi fazi smo ocenili vrednosti kriterija BIC,

v drugi pa izvedli segmentacijo z relativno določenim pragom. V ta namen smo združili dva obstoječa postopka segmentacije. Ocene vrednosti kriterija BIC smo uporabili tudi v drugem predlaganem postopku segmentacije z združevanjem različnih akustičnih predstavitev posnetkov. Tu smo na podlagi ocen vrednosti kriterija BIC izvajali normalizacijo ocen kriterijev različnih predstavitev, ki smo jih potem lahko združevali s postopki fuzije. S fuzijo posameznih predstavitev smo zmanjšali število parametrov modelov, ki jih je bilo potrebno ocenjevati pri segmentaciji s kriteriji BIC, in s tem bolje modelirali krajše odseke segmentacije v primerjavi z modeli skupnih predstavitev.

Razvrščanje segmentov po govorcih je potekalo z združevanjem segmentov. To smo izvajali s postopki hierarhičnega rojenja. Pri tem smo iskali ustrezne predstavitve zvočnih signalov, primerne za združevanje segmentov po govorcih, preizkušali različne mere združevanja segmentov in razvijali kriterije zaustavitve postopkov rojenja. Za predstavitve govornih segmentov smo iskali takšne predstavitve, s katerimi bi bolje opisovali govorčeve lastnosti v segmentih in ne splošnih akustičnih lastnosti segmentov. V prvem pristopu rojenja smo tako segmente modelirali z GMM modeli, ki smo jih izpeljali iz splošnih GMM modelov govora s postopkom MAP adaptacije. V tem primeru smo preizkusili več mer združevanja tako predstavljenih segmentov in predlagali novo mero združevanja s kriterijem BIC, ki pri tem postopku rojenja še ni bila uporabljena. V drugem primeru smo predstavili segmente z združevanjem akustične in prozodične informacije. Prozodične značilke, ki so služile kot dopolnilna informacija osnovnim akustičnim značilkam, smo izpeljali tako, da smo z njimi modelirali osnovne govorčeve prozodične lastnosti. V tem primeru smo morali tudi ustrezno spremeniti postopek rojenja z združevanjem v postopek rojenja s fuzijo dveh predstavitev. Zato smo izvajali normalizacijo ocen mer združevanja in vpeljali skupni kriterij združevanja, ki je temeljil na uteženi vsoti normaliziranih mer združevanja posameznih predstavitev.
V postopkih rojenja smo se ukvarjali tudi s kriteriji zaustavitve rojenja. Predlagali smo dva nova kriterija. Prvi je bil izpeljan iz mere združevanja segmentov s kriterijem BIC, drugi pa na podlagi merjenja napak razvrščanja segmentov med dvema različnima postopkoma združevanja.

Različni preizkusi delovanja in vrednotenja postopkov detekcije govora, segmentacije in razvrščanja segmentov so bili izvedeni na dveh podatkovnih zbirkah zvočnih posnetkov informativnih oddaj. Pri tem smo del posnetkov uporabili za ocenjevanje osnovnih parametrov modelov postopkov, del za razvojne zbirke, s katerimi smo ocenjevali preostale, v glavnem enoštevilske parametre postopkov, in za testne zbirke, ki so služile za vrednotenje in primerjanje postopkov med seboj. Zaradi različnih akustičnih lastnosti posnetkov, ki so bili zajeti v posameznih zbirkah, smo lahko bolj objektivno ocenjevali in primerjali učinkovitost posameznih postopkov v različnih pogojih delovanja.

6.2 Pomen doseženih ciljev

Osnovni namen postopkov, s katerimi smo se ukvarjali v disertaciji, je bilo strukturiranje in organizacija zvočnih posnetkov informativnih oddaj za namene prilagajanja modelov govora v sistemih za razpoznavanje govora in za izgradnjo sistemov indeksacije

zvočnih podatkov. Postopki detekcije govora, samodejne segmentacije in razvrščanja segmentov po govorcih se običajno v takšnih sistemih izvajajo v fazah predobdelave zvočnih posnetkov in so namenjeni osnovni pripravi zvočnih posnetkov za nadaljnjo obdelavo.

S postopki detekcije določamo govorne odseke v zvočnih posnetkih, s postopki segmentacije pa razdelimo govorne odseke na krajše segmente glede na lastnosti, ki jih želimo. Na ta način razdelimo zvočne posnetke na ne–govorne dele in krajše govorne segmente, ki predstavljajo osnovne enote obdelave v sistemih za razpoznavanje govora in indeksacije. Z združevanjem segmentov po govorcih pa povežemo skupaj tiste segmente, ki pripadajo istemu govorcu. Tako pridobimo dodatno informacijo o prisotnosti posameznega govorca v zvočnem posnetku, ki jo uporabimo za prilagajanje govornih modelov glede na govorce v sistemih za razpoznavanje govora oziroma za določanje identifikacije govorcev v sistemih indeksacije zvočnih posnetkov po govorcih. Zato je pravilno in učinkovito delovanje takšnih sistemov močno odvisno od zanesljivega delovanja postopkov, s katerimi smo se ukvarjali v disertaciji.

V nadaljevanju bomo podrobneje opredelili vključevanje predlaganih postopkov v različne sisteme govornih tehnologij, kjer se bomo osredotočili na izgradnjo sistema za indeksacijo zvočnih posnetkov po govorcih.

6.2.1 Vključevanje postopkov v različne sisteme govornih tehnologij

V uvodnem poglavju smo umestili postopke in naloge iz disertacije v področja govornih aplikacij za obdelavo zvočnih posnetkov informativnih oddaj. V tem razdelku pa bomo pregledali možnosti vključevanja teh postopkov v splošne sisteme govornih tehnologij.

Skupna lastnost postopkov, s katerimi smo se ukvarjali v disertaciji, je, da delujejo v več fazah in niso namenjeni sprotni obdelavi zvočnih posnetkov, zato jih lahko vključujemo v sisteme, kjer ni potrebna takšna obdelava posnetkov, ampak je pomemben predvsem zanesljiv rezultat analize zvočnih podatkov. Takšnih sistemov govornih tehnologij je več. V splošnem jih razdelimo na dve skupini: na sisteme, ki temeljijo na razpoznavanju govora in na sisteme, ki temeljijo na razpoznavanju govorcev. V nadaljevanju bomo tako predstavili možnosti vključevanja in uporabo naših postopkov pri obeh skupinah sistemov govornih tehnologij.

Osrednja naloga sistemov za razpoznavanje govora je pretvorba govornih posnetkov v tekstovne prepise, ki jih potem uporabljamo za različne namene in naloge, ki jih rešujemo. Pri tem je potrebno v zvočnih posnetkih najprej določiti govorne odseke. Razpoznavanje ne–govornih delov je namreč nepotrebno, predstavlja pa dodatno delo in povzroča nepravilno delovanje sistemov za razpoznavanje govora. Zato je potrebno v takšnih sistemih najprej zagotoviti zanesljivo detekcijo govora. To dosežemo s postopki, ki smo jih predstavili v poglavju 3. Običajno se v takšnih sistemih izvede najprej segmentacija zvočnih posnetkov na govorne in ne–govorne odseke, nato pa se izvede razpoznavanje govora samo na govornih delih. Postopki detekcije govora so zato praviloma vključeni v prvih fazah obdelave zvočnih posnetkov v sistemih za razpozna-

vanje govora. Rezultat razpoznavanja takšnih sistemov so tako segmenti ne–govora in segmenti govora s tekstovnimi prepisi. Učinkovita detekcija govora nam zagotavlja hitrejše delovanje takšnih sistemov, saj razpoznavanje poteka samo na govornih odsekih. Zato tudi ni potrebno vključevati v sisteme za razpoznavanje govora posebne modele, s katerimi bi modelirali ne–govorne pojave v govoru. Kar pomeni, da so takšni sistemi manj kompleksni in tudi zaradi tega lahko hitreje delujejo. Dodatna prednost detekcije govora je tudi v tem, da zvočne posnetke razbijemo na manjše govorne odseke, ki jih lahko s sistemi za razpoznavanje govora lažje obdelamo. S tem dosežemo boljše delovanje takšnih sistemov.
Po drugi strani lahko v sisteme za razpoznavanje govora vključimo tudi postopke segmentacije in razvrščanja segmentov po govorcih. To je primerno predvsem v sistemih LVCSR, ki so namenjeni razpoznavanju tekočega govora. Segmentacija zvočnih posnetkov na krajše govorne odseke nam v takšnih sistemih omogoča natančnejše in učinkovitejše razpoznavanje govora zaradi manjše količine podatkov, ki jih je potrebno naenkrat obdelati. Dodatno združevanje segmentov glede na govorce pa uporabljamo predvsem v sistemih, kjer se pred razpoznavanjem govora izvajajo postopki prilagajanja govornih modelov glede na govorce *(ang. speaker adaptation)*. Številne študije [Zhang-02, Pusateri-02] so namreč pokazale, da s prilagajanjem govornih modelov na govorce znatno izboljšamo rezultate razpoznavanja.

Velja poudariti, da smo tu imeli v mislih takšne sisteme razpoznavanja govora, ki ne delujejo v realnem času, ampak so namenjeni predvsem kasnejši analizi zvočnih posnetkov. Aplikacije, ki uporabljajo takšne sisteme, so namenjene predvsem pridobivanju različnih informacij iz zvočnih posnetkov informativnih oddaj, kot so strukturiranje in sledenje novicam glede na njihove vsebine, iskanje ključnih besed v zvočnih posnetkih, poravnava zvočnih posnetkov s tekstovnimi prepisi, pridobivanje informacij iz multimedijskih zbirk ipd. V primeru sprotnega razpoznavanja govora bi bilo potrebno postopke, s katerimi smo se ukvarjali v disertaciji, prilagoditi in drugače vključevati v sisteme razpoznavanja govora.

Drugo področje govornih tehnologij, ki vključuje postopke iz disertacije, predstavljajo sistemi, ki temeljijo na razpoznavanju govorcev. Tu imamo v mislih predvsem sisteme iskanja in sledenja govorcev v multimedijskih podatkovnih zbirkah ter sisteme indeksacije zvočnih podatkov glede na govorce za organiziranje zvočnih posnetkov v podatkovne zbirke. V takšnih sistemih predstavljajo segmenti po govorcih osnovne enote analize. Ravno tako je potrebno tudi tu najprej določiti govorne in ne–govorne dele v zvočnih posnetkih, saj se segmentacija praviloma izvaja samo na govornih delih. Združevanje segmentov po govorcih ali po kakšnih drugih lastnostih pa je odvisno od nalog, ki jih rešujemo s takšnimi sistemi. Postopki detekcije govora, segmentacije in razvrščanja segmentov po govorcih, s katerimi smo se ukvarjali v disertaciji, so bili namenjeni predvsem izgradnji sistema indeksacije zvočnih posnetkov po govorcih. Zato bomo njihovo uporabo in vključevanje v tak sistem podrobneje opisali v naslednjem razdelku.

Možna je tudi kombinacija sistemov razpoznavanja govora in razpoznavanja govorcev, ki nam omogoča združevanje in pridobivanje različnih tipov informacij iz zvočnih posnetkov. Z uporabo postopkov iz disertacije tako lahko osnovnim tekstovnim prepisom govora dodamo še informacijo o govorcih, kvaliteti posnetkov, akustičnem ozadju, o

načinu govora ipd. To nam omogoča organiziranje zvočnih posnetkov glede na različne tipe podatkov, ki jih iščemo.

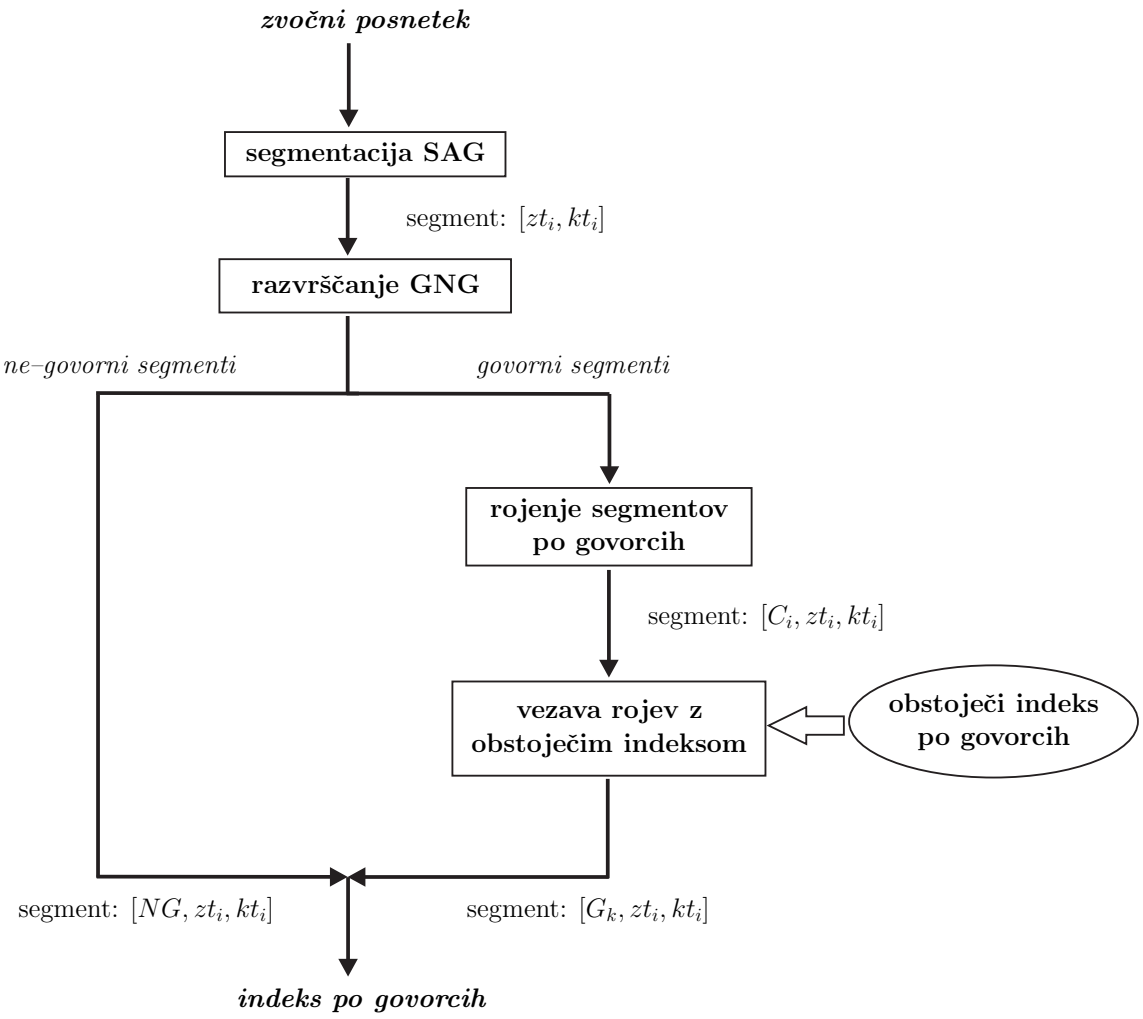
6.2.1.1 Izgradnja sistema za samodejno indeksacijo zvočnih posnetkov

Namen indeksacije v podatkovnih zbirkah je organizirati podatke tako, da lahko izvajamo hitro in učinkovito pridobivanje informacij, ki jih iščemo. V primeru samodejne indeksacije zvočnih posnetkov tako poteka strukturiranje podatkov glede na informacije, ki so zajete v teh posnetkih.

V našem primeru smo si zamislili sistem indeksacije zvočnih posnetkov informativnih oddaj po govorcih. To pomeni, da je potrebno organizirati zvočne posnetke informativnih oddaj v podatkovno zbirko tako, da so primerni za iskanje po posameznih govorcih. Osnovna naloga pri takšni indeksaciji je izgradnja indeksa govorcev, ki je sestavljen iz podatkov o segmentih posameznih govorcev. Segmenti so v tem primeru opremljeni z informacijo, iz katere informativne oddaje smo pridobili segment, s časom trajanja segmenta (začetek in konec) in z oznako (indeksom) govorca. Medtem ko je informacija o informativni oddaji običajno podana že z zvočnim posnetkom, je potrebno vse ostale oznake pridobiti samodejno. Pri tem uporabljamo postopke, s katerimi smo se ukvarjali v disertaciji.

Na sliki 6.1 je prikazana zasnova sistema za indeksacijo zvočnih posnetkov po govorcih. Pri tem smo se zgledovali po podobnih sistemih, ki so bili predstavljeni v [Meignier-02] in [Ajmera-04].

Na sliki 6.1 je prikazano delovanje sistema pri obdelavi zvočnega posnetka ene informativne oddaje. Sistem je zgrajen modularno. V vsaki fazi postopka pridobivamo potrebne informacije za izgradnjo indeksa po govorcih in ga v zadnji fazi pridružimo že obstoječemu indeksu zvočnih posnetkov, ki je shranjen skupaj s podatkovno zbirko. V prvi fazi, pri segmentaciji SAG, zvočni posnetek razdelimo na homogene odseke glede na akustične spremembe in glede na zamenjave govorcev. Tako pridobimo osnovne segmente indeksacije, ki so opremljeni z informacijo o času začetka in konca segmenta (segment: $[zt_i, kt_i]$ na sliki 6.1). V drugem modulu (razvrščanje GNG) se izvaja razvrščanje segmentov na govorne in ne–govorne segmente. Če je segment razvrščen v ne–govorni razred, ga opremimo z oznako NG (segment: $[NG, zt_i, kt_i]$ na sliki 6.1) in ga ne obdelujemo naprej. Izdelava indeksa tako poteka samo na govornih segmentih. Te v nadaljevanju opremimo z oznako pripadnosti segmenta posameznemu roju, ki zajema segmente enega govorca. To dosežemo s postopki rojenja segmentov po govorcih. Segmente tako opremimo z oznakami C_i, ki označujejo posamezne roje segmentov (segment: $[C_i, zt_i, kt_i]$ na sliki 6.1). V zadnjem modulu se izvaja vezava segmentov *(ang. segments tying)* k obstoječim govorcem, ki so zajeti v podatkovni zbirki. V tej fazi je potrebno indeks segmentov, ki smo ga pridobili z rojenjem, uskladiti z obstoječim indeksom govorcev iz podatkovne zbirke. Tu imamo dve možnosti. V primeru, da segmenti roja zajemajo govor govorca, ki je že vsebovan v indeksu govorcev podatkovne zbirke, te segmente označimo z identifikacijo govorca, ki mu pripadajo. V primeru, ko segmenti predstavljajo novega govorca, ki ga v podatkovni zbirki še ni, moramo dodati tega govorca v obstoječi indeks in ustrezno uskladiti oznake govorcev in segmentov. Is-



Slika 6.1: Zasnova sistema za samodejno indeksacijo zvočnih posnetkov po govorcih.

kanje pripadnosti segmentov k obstoječim govorcem se običajno izvaja z GMM modeli s postopki razpoznavanja govorcev [Reynolds-95]. Odločitve, ali segmenti pripadajo obstoječim govorcem ali opisujejo nove govorce, pa se izvajajo na podlagi mer zaupanja *(ang. confidence measures)* v ocene pripadnosti segmentov govorcem v zbirki. V tej fazi postopka segmenti pridobijo oznake G_k (segment: $[G_k, zt_i, kt_i]$ na sliki 6.1), ki so usklajene z oznakami govorcev obstoječega indeksa. Tako označene segmente in nov indeks po govorcih shranimo v podatkovno zbirko. Pri tem omenimo še, da ne–govorne segmente ne vodimo v indeksu govorcev in jih zato tudi ne shranjujemo.

Takšen sistem indeksacije ni zasnovan za sprotno obdelavo zvočnih posnetkov, ampak je potrebno zvočne posnetke najprej pridobiti, sama indeksacija pa se izvaja naknadno. Zato je mogoče v posameznih modulih sistema uporabljati postopke, ki smo jih predlagali v disertaciji. V fazi segmentacije SAG tako lahko uporabimo enega izmed predlaganih postopkov segmentacije po govorcih, ki smo ga opisali v poglavju 4. Ker se izvaja segmentacija v prvi fazi obdelave zvočnih posnetkov, je potrebno izbrati takšen postopek, ki bo neobčutljiv na različne pogoje delovanja. V naših eksperimentih se je za najbolj robustnega izkazal postopek s kriterijem BIC in relativno določenim

pragom. Pri tem je potrebno poudariti, da lahko relativni prag odločitve postavimo tako, da določamo večje število segmentov, kot je dejanskih, saj se v fazi rojenja segmenti, ki pripadajo istemu govorcu, združujejo in tako pridobijo iste oznake. Bistvo postopka segmentacije v tej fazi indeksacije je pridobiti osnovne enote indeksacije, ki jih predstavljajo segmenti posameznih govorcev z enakim akustičnim ozadjem.
V drugi fazi se izvaja razvrščanje na govorne in ne–govorne segmente. Ker izvajamo segmentacijo pred razvrščanjem, je tu primerna uporaba fonetičnih značilk CVS (VUS), ki smo jih predlagali v poglavju 3, z metodo *BICseg–GMM*. Se pravi, da značilke CVS (VUS) računamo na segmentih, ki smo jih pridobili s segmentacijo SAG, razvrščanje pa poteka s pomočjo GMM modelov.
Združevanje segmentov v roje se v nadaljevanju izvaja samo na govornih segmentih. Postopke rojenja, ki jih lahko tu uporabljamo, smo opisali v poglavju 5. V tej fazi je potrebno segmente združiti v roje tako, da vsak roj predstavlja enega govorca. Če predpostavimo, da zvočne posnetke razdelimo na večje število segmentov, kot je dejanskih, potem moramo v fazi združevanja segmentov izbrati tak postopek rojenja, ki bo deloval zanesljivo tudi pri združevanju krajših segmentov. Za to pa se je za najbolj primernega izkazal postopek rojenja segmentov z združevanjem akustične in prozodične informacije, ki ima še eno prednost pred drugimi postopki. Združevanje segmentov poteka namreč tako, da se večji segmenti združujejo pred krajšimi. S tem dosežemo manjšo napako združevanja večjih segmentov, kar je zelo pomembno za indeksacijo, saj se večina podatkov pripiše pravim govorcem, pri manjših (večinoma nepomembnih) segmentih pa je napaka večja. To hkrati pomeni, da tudi s kriteriji zaustavitve ne naredimo večje napake, četudi ne ocenimo pravilno dejanskega števila govorcev v posnetkih, zato bi lahko izbrali med različnimi kriteriji zaustavitve. V našem primeru se je za dokaj učinkovitega izkazal skupni kriterij BIC, s katerim smo dosegli primerljive rezultate z drugim predlaganim kriterijem relativnega DER, kjer je bilo potrebno izvajati dva postopka rojenja.
Zadnja komponenta postopka indeksacije je usklajevanje obstoječega indeksa podatkovne zbirke z indeksom, ki ga pridobimo po rojenju segmentov trenutnega zvočnega posnetka. S postopki iz te faze se v raziskovalnem delu disertacije nismo posebej ukvarjali. Pri izvedbi tega modula bi se zgledovali po standardnih pristopih, ki se uporabljajo pri razpoznavanju govorcev. To pomeni, da bi za vsakega govorca, ki je zajet v podatkovni zbirki, zgradili svoj model. Običajno se v tem primeru uporabljajo GMM modeli v kombinaciji s postopkom MAP adaptacije UBM modelov [Reynolds-95]. V tem primeru bi za podatke iz vsakega roja segmentov ugotavljali, ali pripadajo posameznemu govorcu ali ne. Pri tem bi bilo potrebno podati tudi mero zaupanja v takšno oceno. Na podlagi te mere bi lahko ugotavljali, ali podatki iz posameznega roja pripadajo množici govorcev, ki so zajeti v zbirki, ali predstavljajo nove govorce. V prvem primeru bi dodali segmente govorcu iz zbirke, torej bi obstoječi indeks samo popravili, v drugem pa bi dodali novega govorca v indeks in mu pripisali segmente iz zvočnega posnetka, ki ga obdelujemo. Izvedba vezave rojev z obstoječimi podatki z uporabo GMM in UBM modelov je možna tudi v kombinaciji s postopkom rojenja, ki je temeljil na uporabi metod razpoznavanja govorcev in smo ga opisali v poglavju 5.

Tako pripravljene in označene zvočne posnetke lahko shranjujemo v podatkovne zbirke. Sam postopek iskanja govorcev v takšnih zbirkah je preprost. Na podlagi indeksa, ki ga izdelamo v fazi indeksacije, pridobimo podatke, v katerem zvočnem posnetku in na

katerem mestu imamo shranjen govor govorca, ki ga iščemo. V fazi iskanja na podlagi teh podatkov tvorimo posnetke govora iskanega govorca ter ponudimo seznam povezav na posamezne (osnovne) segmente govora govorca, opremljene z imenom zvočnega posnetka in z informacijo o položaju in trajanju segmenta v posnetku.

Z vključevanjem dodatne informacije v sistem indeksacije bi bilo možno izvajati iskanje tudi po drugih vsebinah. Tako bi lahko v zadnjo fazo obdelave osnovnih segmentov v shemi indeksacije na sliki 6.1 vključili tudi modul za razpoznavanje govora. Na ta način bi pridobili tekstovne prepise govora osnovnih segmentov, ki bi jih lahko uporabljali za nadaljnjo indeksacijo posnetkov po ključnih besedah *(ang. keyword indexing)*, z dodatno obdelavo pa tudi za indeksacijo po vsebinah informativnih oddaj *(ang. topic detection and tracking)*.

6.3 Smernice za nadaljnje delo

S predlaganimi postopki detekcije govora, samodejne segmentacije in razvrščanja segmentov po govorcih smo bistveno izboljšali rezultate v primerjavi s standardnimi postopki, ki se uporabljajo na teh področjih. To je bila predvsem posledica vključevanja informacije višjega reda (fonetične, prozodične značilke) in združevanje z osnovnimi akustičnimi predstavitvami zvočnih posnetkov. S fonetičnimi predstavitvami smo tako izboljšali delovanje postopkov detekcije govora, z dodajanjem prozodične informacije pa postopke razvrščanja segmentov po govorcih. Uporaba fonetičnih in prozodičnih značilk bi bila zato smiselna tudi v postopkih samodejne segmentacije po govorcih. V tem primeru bi bilo potrebno prilagoditi značilke fonetičnih in prozodičnih predstavitev, da bi jih lahko ustrezno ocenjevali tudi na krajših odsekih govora. Ustrezno bi bilo potrebno prilagoditi tudi predlagane postopke segmentacije.

Drugi izzivi za nadaljnje delo zajemajo predvsem vključevanje predlaganih postopkov v različne sisteme govornih tehnologij. V tem poglavju smo že nakazali eno možno izvedbo uporabe postopkov iz disertacije v sistemu za samodejno indeksacijo zvočnih posnetkov po govorcih. Ker so postopki zasnovani tako, da delujejo samostojno, jih lahko vključujemo v sisteme kot samostojne enote ali module. Glede na različne namene uporabe sistemov bi tako lahko ločeno ustrezno prilagajali tudi naše postopke. Osredotočiti pa bi se morali zlasti na zmanjšanje časovne zahtevnosti postopkov in izvedbo postopkov za sprotno delovanje, kar bi povečalo njihovo uporabnost predvsem v sistemih za razpoznavanje govora, ki delujejo v realnem času.

A Preizkusne podatkovne zbirke zvočnih posnetkov

V dodatku so podane podrobnejše razdelitve zvočnih posnetkov informativnih oddaj iz zbirk SiBN in COST278, ki smo jih uporabljali v posameznih preizkusih postopkov iz disertacije. V splošnem je bila izvedena razdelitev zvočnih posnetkov iz obeh zbirk, predstavljenih v poglavju 2, na učne, razvojne in testne množice posnetkov. Učne posnetke smo uporabljali za določitev osnovnih parametrov modelov posameznih postopkov, v glavnem za določitev parametrov GMM in HMM modelov. Razvojni posnetki so bili namenjeni določanju ostalih odprtih parametrov postopkov. To so bile v glavnem enoštevilske vrednosti, ki smo jih določali glede na optimalne rezultate delovanja postopkov na razvojnih posnetkih. Takšni parametri so bili npr. λ pri kriteriju BIC, dolžina oken in preskoka analize pri postopkih segmentacije, uteži v sistemih fuzije, pragovi za meje oziroma za zaustavitev postopkov rojenja ipd. Testni posnetki pa so bili uporabljeni samo za preizkušanje in primerjavo postopkov med seboj.

Razdelitev zvočnih posnetkov na učne, razvojne in testne množice je bila izvedena ročno. Pri tem smo seveda upoštevali osnovno pravilo, da en posnetek lahko pripada samo eni množici. Medtem ko so bile testne množice pri večini preizkusov enake, smo velikost učnih in razvojnih množic prilagajali glede na postopke, ki smo jih preizkušali. V nadaljevanju bomo podali podrobnejše razdelitve zvočnih posnetkov informativnih oddaj, ki smo jih uporabljali pri razvoju in testiranju postopkov detekcije govornih delov, samodejne segmentacije in razvrščanja segmentov po govorcih.

A.1 Detekcija govornih delov v zvočnih posnetkih

Pri razvoju postopkov detekcije govornih delov v zvočnih posnetkih smo uporabljali vse tri množice zvočnih posnetkov.

Učna množica je vsebovala 3 različne posnetke v skupnem trajanju okoli 3 ure. Posnetki so zajemali dve informativni oddaji, eno v slovenskem, drugo v portugalskem jeziku in del zabavne oddaje v slovenskem jeziku. Seznam posnetkov, ki smo jih uporabljali za ocenjevanja GMM modelov govora in ne–govora, je:

```
dnevnik-050603-1900,
spet-doma-190605-16k-part02,
PT_2000_04_20-13_00_00-JornaldaTarde-8.
```

Razvojna množica je bila sestavljena iz približno dveh ur zvočnih posnetkov dveh različnih zabavnih oddaj v slovenskem in italijanskem jeziku. Razvojne posnetke smo uporabljali predvsem za določanje uteži modelov detekcije in uteži fuzije ter nastavitev parametrov HMM mreže za segmentacijo. Posnetka sta bila:

`spet-doma-190605-16k-part01`,
`tv-koper-16k`.

Prvi posnetek je bil del zabavne oddaje, ki se predvaja ob nedeljah zvečer na RTV Slovenija, druga pa je bila informativno–zabavna oddaja televizijske postaje TV Koper-Capodistria. Skupna lastnost oddaj je, da vsebujejo veliko ne–govornih odsekov, zato sta bili primerni za ocenjevanje skupne mere detekcije govornih in ne–govornih odsekov.

Testni posnetki so bili sestavljeni iz posnetkov iz zbirk SiBN in COST278. V primeru SiBN zbirke smo izbrali vse posnetke iz zbirke, razen posnetka `dnevnik-050603-1900`, ki smo ga uporabili že v učni množici. Vseh testnih posnetkov iz zbirke SiBN je bilo 33. Seznam posnetkov je bil naslednji:

`dnevnik-090603-1900`,
`dnevnik-090703-1900`,
`dnevnik-100603-1900`,
`dnevnik-100703-1900`,
`dnevnik-110603-1900`,
`dnevnik-110703-1900`,
`dnevnik-130603-1900`,
`dnevnik-140703-1900`,
`dnevnik-150503-1900`,
`dnevnik-150703-1900`,
`dnevnik-160603-1900`,
`dnevnik-160703-1900`,
`dnevnik-170603-1900`,
`dnevnik-180603-1900`,
`dnevnik-190603-1900`,
`dnevnik-190803-1900`,
`dnevnik-200503-1900`,
`dnevnik-200603-1900`,
`dnevnik-200803-1900`,
`dnevnik-220703-1900`,
`dnevnik-230603-1900`,
`dnevnik-230703-1900`,
`dnevnik-240603-1900`,
`dnevnik-240703-1900`,
`dnevnik-250703-1900`,
`dnevnik-260603-1900`,
`dnevnik-260803-1900`,
`dnevnik-270603-1900`,
`dnevnik-270803-1900`,

dnevnik-280703-1900,
dnevnik-290703-1900,
dnevnik-300703-1900,
dnevnik-310703-1900.

Pri zbirki COST278 smo prav tako uporabili vse posnetke iz zbirke, razen posnetkov PT_2000_04_20-13_00_00-JornaldaTarde-8 in SI_dnevnik-050603-1900, ki sta bila uporabljena v učni množici. Vseh testnih posnetkov v tem primeru je bilo 55. Seznam posnetkov je bil naslednji:

BE_journaal_20000614,
BE_journaal_20000921,
BE_journaal_20010326,
BE_journaal_20010619,
BE_journaal_20010701,
BE_journaallaat_20010716,
CZ_03_06_10_CT1_udalosi,
CZ_03_06_12_CT1_udalosi,
CZ_03_06_15_CT1_udalosi,
CZ_03_06_16_CT1_udalosi,
CZ_04_10_07_CT1_udalosti_komentare,
CZ_04_10_13_CT1_vecernik_z_cech,
CZ_04_10_20_CT1_udalosti_komentare,
GA_M021009,
GA_M021010,
GA_M021011,
GR_NET_20010805_2000_Daily_News_16k,
GR_NET_20010808_2000_Daily_News_16k,
GR_NET_20010928_2000_Daily_News_16k,
HR_HRTdnevnik031003,
HR_HRTdnevnik041003,
HR_HRTdnevnik051003,
HR_HRTdnevnik061003,
HR_HRTdnevnik071003,
HR_HRTdnevnik101003,
HU_MTV1_2004_0220_1200,
HU_MTV1_2004_0227_0800,
HU_MTV1_2004_0227_0812,
HU_RTL_2004_0226_1200,
HU_RTL_2004_0227_1200,
HU_RTL_2004_0302_1200,
HU_RTL_2004_0303_0800,
HU_RTL_2004_0304_1200,
HU_TV2_2004_0226_1830,
HU_TV2_2004_0303_1830,
HU_TV2_2004_0304_1830,
PT_2000_04_06-20_00_00-Telejornal-1,

PT_2000_04_17-20_30_00-RTPEconomia-3,
PT_2000_04_18-19_30_00-PaisRegioesLx-5,
PT_2000_04_18-21_00_00-Jornal2-2,
PT_2000_07_19-00_00_00-24Horas-11,
SI2_TVS-odmevi-2002-03-25-2200,
SI2_TVS-odmevi-2002-09-18-2200,
SI2_TVS-odmevi-2002-09-30-2200,
SI_dnevnik-150503-1900,
SI_dnevnik-200503-1900,
SK_TA3-2003-05-29_07-59-59h_16kHz,
SK_TA3-2003-05-29_11-59-58h_16kHz,
SK_TA3-2003-05-30_09-00-00h_16kHz,
SK_TA3-2003-05-30_21-19-04h_16kHz,
SK_TA3-2003-05-31_22-29-59h_16kHz,
SK_TA3-2003-06-01_22-29-59h_16kHz,
SK_TA3-2003-06-11_22-59-46h_16kHz,
SK_TA3-2003-06-12_12-59-46h_16kHz,
SK_TA3-2003-06-12_23-00-19h_16kHz.

A.2 Samodejna segmentacija zvočnih posnetkov

Pri razvoju postopkov segmentacije zvočnih posnetkov glede na zamenjave govorcev in glede na spremembe akustičnih ozadij smo potrebovali samo razvojne in testne množice posnetkov. Učnih posnetkov, ki bi jih uporabljali za določanje osnovnih parametrov modelov, kot je bilo to v primeru detekcije govora, nismo potrebovali, ker takšnih modelov v naših postopkih segmentacije nismo uporabljali.

Določali pa smo različne pragove odločitve za mejo med dvema segmentoma, uteži fuzije pri postopkih segmentacije z združevanjem različnih predstavitev ter številne ostale odprte parametre različnih postopkov segmentacije, kar smo podrobneje predstavili v poglavju 4. Za to smo potrebovali razvojno množico posnetkov, ki smo jo sestavili iz sedmih posnetkov iz zbirke SiBN. Seznam posnetkov je bil naslednji:

dnevnik-050603-1900,
dnevnik-150503-1900,
dnevnik-160603-1900,
dnevnik-190603-1900,
dnevnik-200803-1900,
dnevnik-270803-1900,
dnevnik-310703-1900.

Med razvojne posnetke nismo vključili nobenega posnetka iz zbirke COST278, saj smo na ta način želeli preizkušati neobčutljivost delovanja postopkov na različne akustične pogoje, ki so bili zajeti v obeh zbirkah.

Testne posnetke smo pridobili iz preostanka posnetkov iz zbirke SiBN in celotne zbirke

COST278. V primeru SiBN nam je tako ostalo 27 posnetkov, ki so zbrani v naslednjem seznamu:

dnevnik-090603-1900,
dnevnik-090703-1900,
dnevnik-100603-1900,
dnevnik-100703-1900,
dnevnik-110603-1900,
dnevnik-110703-1900,
dnevnik-130603-1900,
dnevnik-140703-1900,
dnevnik-150703-1900,
dnevnik-160703-1900,
dnevnik-170603-1900,
dnevnik-180603-1900,
dnevnik-190803-1900,
dnevnik-200503-1900,
dnevnik-200603-1900,
dnevnik-220703-1900,
dnevnik-230603-1900,
dnevnik-230703-1900,
dnevnik-240603-1900,
dnevnik-240703-1900,
dnevnik-250703-1900,
dnevnik-260603-1900,
dnevnik-260803-1900,
dnevnik-270603-1900,
dnevnik-280703-1900,
dnevnik-290703-1900,
dnevnik-300703-1900.

V primeru zbirke COST278 smo za testiranje postopkov segmentacije uporabili vse posnetke, razen posnetka SI_dnevnik-050603-1900, ki je bil uporabljen v razvojni zbirki. Vseh posnetkov je bilo tako 56 in so zbrani v naslednjem seznamu:

BE_journaal_20000614,
BE_journaal_20000921,
BE_journaal_20010326,
BE_journaal_20010619,
BE_journaal_20010701,
BE_journaallaat_20010716,
CZ_03_06_10_CT1_udalosi,
CZ_03_06_12_CT1_udalosi,
CZ_03_06_15_CT1_udalosi,
CZ_03_06_16_CT1_udalosi,
CZ_04_10_07_CT1_udalosti_komentare,
CZ_04_10_13_CT1_vecernik_z_cech,

CZ_04_10_20_CT1_udalosti_komentare,
GA_M021009,
GA_M021010,
GA_M021011,
GR_NET_20010805_2000_Daily_News_16k,
GR_NET_20010808_2000_Daily_News_16k,
GR_NET_20010928_2000_Daily_News_16k,
HR_HRTdnevnik031003,
HR_HRTdnevnik041003,
HR_HRTdnevnik051003,
HR_HRTdnevnik061003,
HR_HRTdnevnik071003,
HR_HRTdnevnik101003,
HU_MTV1_2004_0220_1200,
HU_MTV1_2004_0227_0800,
HU_MTV1_2004_0227_0812,
HU_RTL_2004_0226_1200,
HU_RTL_2004_0227_1200,
HU_RTL_2004_0302_1200,
HU_RTL_2004_0303_0800,
HU_RTL_2004_0304_1200,
HU_TV2_2004_0226_1830,
HU_TV2_2004_0303_1830,
HU_TV2_2004_0304_1830,
PT_2000_04_06-20_00_00-Telejornal-1,
PT_2000_04_17-20_30_00-RTPEconomia-3,
PT_2000_04_18-19_30_00-PaisRegioesLx-5,
PT_2000_04_18-21_00_00-Jornal2-2,
PT_2000_04_20-13_00_00-JornaldaTarde-8,
PT_2000_07_19-00_00_00-24Horas-11,
SI2_TVS-odmevi-2002-03-25-2200,
SI2_TVS-odmevi-2002-09-18-2200,
SI2_TVS-odmevi-2002-09-30-2200,
SI_dnevnik-150503-1900,
SI_dnevnik-200503-1900,
SK_TA3-2003-05-29_07-59-59h_16kHz,
SK_TA3-2003-05-29_11-59-58h_16kHz,
SK_TA3-2003-05-30_09-00-00h_16kHz,
SK_TA3-2003-05-30_21-19-04h_16kHz,
SK_TA3-2003-05-31_22-29-59h_16kHz,
SK_TA3-2003-06-01_22-29-59h_16kHz,
SK_TA3-2003-06-11_22-59-46h_16kHz,
SK_TA3-2003-06-12_12-59-46h_16kHz,
SK_TA3-2003-06-12_23-00-19h_16kHz.

Ker smo v razvojno množico vključili samo posnetke iz zbirke SiBN, smo v primeru testne množice SiBN posnetkov ocenjevali delovanje postopkov segmentacije v razme-

roma podobnih akustičnih razmerah, v primeru testnih posnetkov iz zbirke COST278 pa v različnih akustičnih razmerah.

A.3 Razvrščanje segmentov po govorcih s postopki rojenja

Podobno kot v primeru segmentacije zvočnih posnetkov smo tudi pri razvrščanju segmentov po govorcih uporabljali samo dve množici posnetkov: razvojno in testno. Učne množice nismo načrtovali, ker tudi pri rojenju ni bilo potrebno vnaprej ocenjevati osnovnih parametrov modelov. Možnost uporabe GMM modelov je bila le v primeru rojenja z metodami razpoznavanja govorcev, kjer je bilo potrebno vnaprej pridobiti UBM modele govora, vendar smo v našem primeru UBM modele ocenjevali neposredno iz zvočnih posnetkov, ki smo jih obdelovali.

Razvojna in obe testni množici so bile enake kot v primeru segmentacije. Razlog zato je bil predvsem v tem, da smo v primeru preizkusov samodejne segmentacije želeli ocenjevati delovanje postopkov razvrščanja v kombinaciji s segmentacijo. Razvojne posnetke smo uporabili za določitev vseh odprtih parametrov postopkov, predvsem kriterijev združevanja in kriterijev zaustavitve rojenja. Uporabili smo jih tudi za določitev GMM modelov za razvrščanje segmentov po spolu, ki smo jih potrebovali za določitev UBM modelov v primeru rojenja z metodami razpoznavanja govorcev. Parametri so bili določeni glede na optimalne povprečne rezultate mere DER na vseh sedmih posnetkih razvojne množice.

Podobno kot v primeru segmentacije so bili tudi tu zaradi izbire posnetkov v razvojni zbirki postopki rojenja bolj primerni za razvrščanje segmentov zbirke SiBN. S testnimi posnetki iz zbirke COST278 pa smo želeli predvsem oceniti delovanje postopkov razvrščanja v primeru neujemanja akustičnih razmer med razvojnimi in testnimi posnetki.

Viri in literatura

[Ajmera-03] J. Ajmera, I. McCowan & H. Bourlard. *Speech/music segmentation using entropy and dynamism features in a HMM classification framework.* Speech Communication, vol. 40, št. 3, str. 351–363, maj 2003.

[Ajmera-04] J. Ajmera. Robust audio segmentation. doktorska disertacija, EPFL Lausanne, Švica, 2004.

[Aurix-03] Aurix. *Aurix aligner*, 2003. `http://www.aurix.com/product/index.htm`.

[Baker-05] B. Baker, R. Vogt & S. Sridharan. *Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification.* Proceedings of Interspeech 2005 - Eurospeech, Lizbona, Portugalska, september 2005.

[Barras-01] C. Barras, E. Geoffrois, Z. Wu & M. Liberman. *Transcriber: Development and use of a tool for assisting speech corpora production.* Speech Communication, vol. 33, št. 1, str. 5–22, 2001.

[Barras-04] C. Barras, X. Zhu, S. Meignier & J. Gauvain. *Improving Speaker Diarization.* Proceedings of DARPA RT04, Palisades NY, ZDA, november 2004.

[Ben-02] M. Ben, R. Blouet & F. Bimbot. *A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distance.* Proceedings of the ICASSP2002, vol. 1, 2002. Orlando, ZDA.

[Ben-04] M. Ben, M. Betser, F. Bimbot & G. Gravier. *Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs.* Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004), Jeju Island, Koreja, oktober 2004.

[Beyerlein-02] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz & A. Sixtus. *Large vocabulary continuous speech recognition of Broadcast*

News - The Philips/RWTH approach. Speech Communication, vol. 37, št. 1-2, str. 109–131, maj 2002.

[Biatov-03] K. Biatov. *Large Text and Audio Data Alignment for Multimedia Applications.* Václav Matousek & Pavel Mautner, (ur.), Proceedings of the 6th International Conference on Text, Speech and Dialogue, TSD 2003, Lecture Notes in Computer Science 2807, str. 349–356. Springer, 2003. Ceské Budejovice, Češka.

[Bimbot-93] F. Bimbot & L. Mathan. *Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure.* Proceedings of European Conference on Speech Communication and Technology (Eurospeech'93), vol. 1, str. 169–172, september 1993. Berlin, Nemčija.

[BN-DWK-99] DWK. *Deutsche Welle Kalenderblatt*, 1999. `http://www.kalenderblatt.de`.

[Bonastre-00] J.F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin & C. Wellekens. *A speacker tracking system based on speaker turn detection for NIST evaluations.* Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing, ICASSP 2000, str. 1177–1180, 2000.

[Brown-99] R. D. Brown, T. Pierce, Y. Yang & J. G. Carbonell. *Link Detection - Results and Analysis.* Proceedings of the TDT-1999 Workshop, 1999.

[Burnham-03] K. P. Burnham & D. R. Anderson. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer, 2. izdaja, december 2003.

[Calinski-74] R. B. Calinski & J. Harabasz. *A dendrite method for cluster analysis.* Communications in statistics, vol. 3, str. 1–27, 1974.

[Campbell-97] J. G. Campbell, C. Fraley, F. Murtagh & A. E. Raftery. *Linear flaw detection in woven textiles using model–based clustering.* Pattern Recognition Letters, vol. 18, str. 1539–1548, december 1997.

[Carey-99] M. J. Carey, E. S. Parris & H. Lloyd-Thomas. *A comparison of features for speech, music discrimination.* Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999), vol. 2, str. 149–152, Phoenix, Arizona, ZDA, marec 1999.

[Cettolo-00] M. Cettolo & M. Federico. *Model selection criteria for acoustic segmentation.* Proceedings of the ISCA ITRW ASR2000

Automatic Speech Recognition, str. 221–227, Pariz, Francija, 2000.

[Cettolo-05] M. Cettolo, M. Vescovi & R. Rizzi. *Evaluation of BIC-based algorithms for audio segmentation.* Computer Speech and Language, vol. 19, št. 2, str. 147–170, april 2005.

[Chen-98] S. S. Chen & P. S. Gopalakrishnan. *Speaker, environment and channel change detection and clustering via the Bayesian information criterion.* Proceedings of the DARPA Speech Recognition Workshop, 1998.

[Chen-02] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky & P. Olsen. *Automatic transcription of Broadcast News.* Speech Communication, vol. 37, št. 1-2, str. 69–87, maj 2002.

[Dampster-77] A. Dampster, N. Laird & D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, vol. 39, št. 1, str. 1–38, 1977.

[Dasgupta-98] A. Dasgupta & A. E. Raftery. *Detecting features in spatial point processes with clutter via model–based clustering.* Journal of the American Statistical Association, vol. 93, št. 441, str. 294–302, marec 1998.

[Delacourt-01] P. Delacourt & C. J. Wellekens. *DISTBIC: A speaker-based segmentation for audio data indexing.* Speech Communication, vol. 32, št. 1-2, str. 111–126, september 2001.

[Do-03] M. N. Do. *Fast Approximation of Kullback–Leibler Distance for Dependence Trees and Hidden Markov Models.* Signal Processing Letters, vol. 10, str. 115–118, 2003.

[Dudoit-02] S. Dudoit & J. Fridlyand. *A predictioon–based resampling method for estimating the number of clusters in a dataset.* Genome Biology, vol. 3, št. 7, junij 2002.

[Federico-00] M. Federico. *A Baseline System for the Retrieval of Italian Broadcast News.* Speech Communication, Special Issue on Accessing Information in Spoken Audio, vol. 32, str. 37–47, 2000.

[Fiscus-05] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot & C. Laprun. *The Rich Transcription 2005 Spring Meeting Recognition Evaluation.* Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, Velika Britanija, julij 2005. Lecture Notes in Computer Science, Springer Berlin/Heidelberg.

[Fraley-98] C. Fraley & A. E. Raftery. *How many clusters? Which clustering method? Answers via model–based cluster analysis.* The Computer Journal, vol. 41, št. 8, str. 578–588, 1998.

[Furui-98] S. Furui, K. Takagi, A. Iwasaki, K. Ohtsuki, T. Matsuoka & S. Matsunaga. *Japanese Broadcast News Transcription and Topic Detection.* Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, str. 144–149, Lansdowne Conference Resort Lansdowne, Virginia, ZDA, februar 1998.

[Gales-96] M. J. F. Gales & P.C. Woodland. *Mean and Variance Adaptation Within the MLLR Framework.* Computer, Speech & Language, vol. 10, str. 249–264, 1996.

[Galliano-05] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. F. Bonastre & G. Gravier. *The ESTER phase II evaluation campaign for the rich transcription of french broadcast news.* Proceedings of Interspeech 2005 - Eurospeech, str. 1149–1152, Lizbona, Portugalska, september 2005.

[Gallwitz-02] F. Gallwitz, H. Niemann, E. Nöth & V. Warnke. *Integrated recognition of words and prosodic phrase boundaries.* Speech Communication, vol. 36, št. 1–2, str. 81–95, januar 2002.

[Garofolo-93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett & N. L. Dahlgren. *DARPA TIMIT acoustic-phonetic continuous speech corpus.* U.S. Department of Commerce, NIST, Gaithersburg, MD, ZDA, februar 1993.

[Gauvain-00] J.L. Gauvain, L. Lamel, C. Barras, G. Adda & Y. Kercadio. *The LIMSI SDR system for TREC-9.* Proceedings of the Text Retrieval Conference, str. 335–341, november 2000. Gaithersburg.

[Gauvain-02] J. L. Gauvain, L. Lamel & G. Adda. *The LIMSI Broadcast News transcription system.* Speech Communication, vol. 37, št. 1-2, str. 89–108, maj 2002.

[Gauvain-03] J.-L. Gauvain & L. Lamel. *Structuring Broadcast Audio for Information Access.* EURASIP journal on Applied Signal Processing, vol. 2, str. 140–150, 2003.

[Gish-91] H. Gish, M. H. Siu & R. Rohlicek. *Segregation of speakers for speech recognition and speaker identification.* Proceedings of the ICASSP 1991, str. 873–876, Toronto, Kanada, 1991.

[Gordon-99] A. Gordon. Classification. Wiley, New York, ZDA, 2nd izdaja, 1999.

[Graff-02] D. Graff. *An overview of Broadcast News corpora.* Speech Communication, vol. 37, št. 1-2, str. 15–26, maj 2002.

[Greenberg-95] S. Greenberg. *The ears have it: The auditory basis of speech perceptions.* International Congress of Phonetic Sciences (ICPhS 95), str. 34–41, Stockholm, Švedska, avgust 1995.

[Gros-05] J. Ž. Gros, F. Mihelič, T. Erjavec & Š. Vintar. *The VoiceTRAN Speech-to-Speech Communicator.* Proceedings of the 8th International Conference on Text, Speech and Dialogue, TDS 2005, vol. 1, str. 379–384, Karlovy Vary, Češka, september 2005. Springer LNAI 3658.

[Guo-03] G. Guo & S. Z. Li. *Content-Based Audio Classification and Retreival by Support Vector Machines.* IEEE Transactions on Neural Networks, vol. 14, št. 1, str. 209–215, januar 2003.

[Hain-98] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland & S. J. Young. *Segment Generation and Clustering in the HTK Broadcast News Transcription System.* Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop, str. 133–137, Lansdowne, VA, ZDA, februar 1998.

[Hartigan-85] J. A. Hartigan. *Statistical theory in clustering.* Journal of Classification, vol. 2, str. 63–76, 1985.

[Hermansky-90] H. Hermansky. *Perceptual linear predictive (PLP) analysis of speech.* Journal of the Acoustical Society of America, vol. 87, št. 4, str. 1738–1752, 1990.

[Hori-02] C. Hori, S. Furui, R. Malkin, H. Yu & Alex Waibel. *Automatic speech summarization applied to English broadcast news speech.* Proceedings of the ICASSP2002, vol. 1, str. 9–12, 2002. Orlando, ZDA.

[Istrate-05] D. Istrate, N. Scheffer, C. Fredouille & J.-F. Bonastre. *Broadcast News Speaker Tracking for ESTER 2005 Campaign.* Proceedings of Interspeech 2005 - Eurospeech, str. 2445–2448, Lizbona, Portugalska, september 2005.

[Iurgel-01] U. Iurgel, R. Meermeier, S. Eickeler & G. Rigoll. *New Approaches to Audio-Visual Segmentation of TV News for Automatic Topic Retrieval.* Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing, ICASSP 2001, maj 2001. Salt Lake City , Utah, ZDA.

[Iyengar-00] G. Iyengar & C. Neti. *Speaker change detection using joint audio-visual statistics.* Proceedings of the RIAO'2000, Pariz, Francija, april 2000. Paris, France.

[Jain-05] A. Jain, K. Nandakumar & A. Ross. *Score normalization in multimodal biometric systems.* Pattern Recognition, vol. 38, št. 12, str. 2270–2285, december 2005.

[Jin-97] H. Jin, F. Kubala & R. Scwartz. *Automatic speaker clustering.* Proceedings of Speech Recognition Workshop, ARPA, vol. 3, str. 108–111, Chantilly, VA, ZDA, februar 1997.

[Johnson-98] S. Johnson & P. Woodland. *Speaker clustering using direct maximization of the MLLR–adapted likelihood.* International Conference on Spoken Language Processing (ICSLP 1998), vol. 5, str. 1775–1778, Sydney, Avstralija, april 1998.

[Johnson-99] S. Johnson. *Who spoke when? - automatic segmentation and clustering for determining speaker turns.* Proceedings of the Eurospeech'99, vol. 5, str. 2211–2214, september 1999. Budimpešta, Madžarska.

[Johnson-01] S.E. Johnson, P. Jourlin, K Spärck Jones & P.C. Woodland. *Information Retrieval from Unsegmented Broadcast News Audio.* International Journal of Speech Technology (IJST), vol. 4, št. 3-4, str. 251–268, julij 2001.

[JTSI-03] V. Gorjanc (ur.). *Posebna številka: Jezikovne tehnologije za slovenščino.* Jezik in slovstvo, vol. 48, št. 3–4, maj - avgust 2003.

[Kajarekar-03] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke & R. R. Gadde. *Speaker Recognition Using Prosodic and Lexical Features.* Proceedings of IEEE Speech Recognition and Understanding Workshop (ASRU 2003), St. Thomas, Deviški otoki, december 2003.

[Karnebäck-02] S. Karnebäck. *Expanded examinations of a low frequency modulation feature for speech/music discrimination.* International Conference on Spoken Language Processing (ICSLP2002 -Interspeech 2002), str. 2009–2012, Denver, Colorado, ZDA, september 2002.

[Kaufman-99] L. Kaufman & P. J. Rousseeuw. Finding gropus in data: An introduction to cluster analysis. Wiley, New York, ZDA, 1999.

[Kačič-00] Z. Kačič, B. Horvat & A. Zögling. *Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language.* Proceedings of the LREC 2000, Atene, Grčija, 2000.

[Kemp-98] T. Kemp, P. Geutner, M. Schhmidt, B. Tomaz, M. Weber, M. Westphal & A. Waibel. *The Interactive Systems Labs*

View4You video indexing system. Proceedings ICSLP'98, vol. 4, str. 1639–1642, december 1998. Sydney, Avstralija.

[Kemp-00] T. Kemp, M. Schmidt, M. Westphal & A. Waibel. *Strategies for Automatic Segmentation of Audio Data.* Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing, ICASSP 2000, vol. 3, str. 1423–1426, 2000.

[Kozak-97] J. Kozak. Podatkovne strukture in algoritmi. DMFA: Društvo matematikov, fizikov in astronomov Slovenije, Fakulteta za matematiko in fiziko, Univerza v Ljubljani, Slovenija, 1997.

[Krzanowski-85] W. J. Krzanowski & Y. T. Lai. *A criterion for determining the number of groups in a dataset using sum of squares clustering.* Biometrika, vol. 44, str. 23–34, 1985.

[Kubala-97] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz & J. Makhoul. *The 1996 BBN Byblos Hub-4 Transcription System.* Proceedings of the Speech Recognition Workshop, str. 90–93, 1997.

[Lamel-02] L. Lamel & J.L. Gauvain. *Automatic Processing of Broadcast Audio in Multiple Languages.* Proceedings of the Eusipco 2002, september 2002.

[Lapidot-03] I. Lapidot. *SOM as likelihood estimator for speaker clustering.* Proceedings of the Eurospeech 2003, str. 3001–3004, 2003. Ženeva, Švica.

[LDC-00] Linguistic Data Consortium. *Hub-4 Broadcast News Transcription Conventions*, 2000.
`http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/conventons.html`.

[Lee-89] K. F. Lee & H. W. Hon. *Speaker-Independent Phone Recognition Using Hidden Markov Models.* IEEE Transactions on Acoustic Speech and Signal Processing, vol. 37, št. 11, str. 1641–1648, 1989.

[Leek-00] T. Leek, H. Jin, S. Sista & R. Schwartz. *The BBN Crosslingual Topic Detection and tracking System.* Working Notes of the the Third TDT-3 Workshop, 2000.

[Leroux-92] M. Leroux. *Consistent estimation of a mixing distribution.* The Annals of Statistics, vol. 20, str. 1350–1360, 1992.

[Li-02] W. Li, P. Bernaola-Galvan, F. Haghighi & I. Grosse. *Applications of recursive segmentation to the analysis of DNA*

sequences. Computers and Chemistry, vol. 26, str. 491–510, 2002.

[Liu-99] D. Liu & F. Kubala. *Fast speaker change detection for broadcast news transcription and indexing.* Proceedings of the 6th European Conference on Speech Communication and Technology, str. 1031–1034, 1999. Budimpešta, Madžarska.

[Logan-00] B. Logan. *Mel-Frequency Cepstral Coefficients for Music Modeling.* Proceedings of the International Symposium on Music Information Retrieval, Plymouth, Massachusetts, ZDA, oktober 2000.

[Lopez-00] J. F. Lopez & D. P. W. Ellis. *Using acoustic condition clustering to improve acoustic change detection on broadcast news.* Proceedings of the International Conference on Spoken Language Processing, ICSLP 2000, Peking, Kitajska, oktober 2000.

[Lu-02] L. Lu, H.-J. Zhang & S. Z. Li. *Content-based audio classification and segmentation by using support vector machines.* Multimedia Systems, vol. 8, št. 6, str. 482–492, 2002.

[Macherey-02] W. Macherey & H. Ney. *Toward automatic corpus preparation for a German broadcast news transcription system.* Proceednigs of International Conference on Spoken Language Processing (ICSLP 2002), vol. 1, str. 733–736, Denver, CO, ZDA, maj 2002.

[Magrin-Chagnolleau-02] I. Magrin-Chagnolleau & N. Parlangeau-Vallès. *Audio indexing: what has been accomplished and the road ahead.* Proceedings of Joint Conference on Information Sciences, (JCIS 2002), str. 911–914, Durham, North Carolina, ZDA, marec 2002.

[Magrin-99] I. Magrin-Chagnolleau, A. E. Rosenberg & S. Parthasarathy. *Detection of Target Speakers in Audio Databases.* Proceedings of the ICASSP 99, vol. 2, str. 821–824, marec 1999. Phoenix, ZDA.

[Makhoul-00] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz & A. Srivastava. *Speech and Language Technologies for Audio Indexing and Retrieval.* Proceedings of the IEEE 88, vol. 8, str. 1338–1353, avgust 2000.

[Manos-97] A. Manos & V. Zue. *A segment-based wordspotter using phonetic filler models.* Proceedings of the ICASSP 97, vol. 2, str. 899–902, 1997. München, Nemčija.

[Martin-00] A. Martin & M. Przybocki. *The NIST 1999 Speaker Recognition Evaluation–An Overview.* Digital Signal Processing, vol. 10, št. 1–3, str. 1–18, januar 2000.

[Maybury-99] M. Maybury. *Intelligent multimedia for the new milennium.* Proceedings of the Eurospeech'99, vol. 5, str. 1–15, september 1999. Budimpešta, Madžarska.

[Meignier-02] S. Meignier, J. Bonastre & I. Magrin-Chagnolleau. *Speaker utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases.* Proceedings of the ICSLP 2002, september 2002. Denver, Kolorado, ZDA.

[Meinedo-01] H. Meinedo, N. Souto & J. Neto. *Speech recognition of broadcast news for the european portuguese language.* Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU 2001), Madonna di Campiglio, Trento, Italija, december 2001.

[Meinedo-03a] H. Meinedo, D. Caseiro, J. Neto & I. Trancoso. *AUDIMUS.media: A Broadcast News Speech Recognition System for the European Portuguese Language.* Proceedings of the PROPOR'2003, junij 2003.

[Meinedo-03b] H. Meinedo & J. Neto. *Audio segmentation, classification and clustering in a broadcast news task.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, Kitajska, april 2003.

[Mihelič-03] F. Mihelič, J. Gros, S. Dobrišek, J. Žibert & N. Pavešić. *Spoken language resources at LUKS of the University of Ljubljana.* International Journal of Speech Technology (IJST), vol. 6, št. 3, str. 221–232, 2003.

[Moraru-03a] D. Moraru, S. Meignier, L. Besacier, J. Bonastre & Y. Magrin-Chagnolleau. *The Elisa Consortium Approaches In Speaker Segmentation During The Nist 2002 Speaker Recognition Evaluation.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, Kitajska, april 2003.

[Moraru-03b] D. Moraru, S. Meignier, C. Fredouill & J. F. Bonastre. *ELISA, CLIPS and LIA NIST 2003 segmentation.* Proceedings of Spring 2003 Rich Transcription Workshop (RT-03s), 2003. Boston, MA, ZDA.

[Moraru-05] D. Moraru, M. Ben & G. Gravier. *Experiments on speaker tracking and segmentation in radio broadcast news.* Proceedings of Interspeech 2005 - Eurospeech, str. 3049–3052, Lizbona, Portugalska, september 2005.

[Moreno-98] P. Moreno, C. Joerg, J. M. Van Thong & O. Glickman. *A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments.* Proceedings of the International Conference on Spoken Language Processing (ICSLP 1998), december 1998. Sydney, Avstralija.

[Mori-01] K. Mori & S. Nakagawa. *Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition.* Proceedings of ICASSP 2001, vol. 1, str. 413–416, Salt Lake City, Utah, ZDA, 2001.

[MUSA-02] MUSA. *MUSA: Multilingual Subtitling of Multimedia Content*, 2002.
`http://www.esat.kuleuven.ac.be/~spch/cgi-bin/projects.cgi?MUSA`.

[Neti-00] C. Neti, B. Maison, A. Senior, G. Iyengar, P. de Cuetos, S. Basu & A. Verma. *Joint processing of audio and visual information for multimedia indexing and human-computer interaction.* Proceedings of the RIAO'2000, april 2000. Pariz, Francija.

[Nguyen-03] P. Nguyen & J. C. Junqua. *PSTL's Speaker Diarization.* Proceedings of Spring 2003 Rich Transcription Workshop (RT-03s), Boston, MA, ZDA, 2003.

[Nöth-02] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt & H. Niemann. *On the use of prosody in automatic dialogue understanding.* Speech Communication, vol. 36, št. 1–2, str. 45–62, januar 2002.

[Olive-00] LE-4 Olive project. *A Multilingual Indexing Tool for Broadcast Material Based on Speech Recognition*, 2000.
`http://dis.tpd.tno.nl/olive/`.

[Pallett-99] D. S. Pallett, J. G. Fiscus, J. S. Garofolo, A. Martin & M. Przybocki. *1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures.* Proceedings of 1999 DARPA Broadcast News Workshop, 1999.

[Pallett-02] D. S. Pallett. *The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program.* Speech Communication, vol. 37, št. 1-2, str. 69–87, maj 2002.

[Pavešić-00] N. Pavešić. Razpoznavanje vzorcev: Uvod v analizo in razumevanje vidnih in slušnih signalov - 2. razširjena izdaja. Fakulteta za elektrotehniko, Univerza v Ljubljani, Ljubljana, Slovenija, 2000.

[Pelecanos-01] J. Pelecanos & S. Sridharan. *Feature warping for robust speaker verification.* Proceedings of ISCA Workshop on Speaker Recognition - 2001: A Speaker Odyssey, Kreta, Grčija, junij 2001.

[Picone-93] J. W. Picone. *Signal modeling techniques in speech recognition.* Proceedings of the IEEE, vol. 81, št. 9, str. 1215–1247, 1993.

[Potamianos-04] G. Potamianos, C. Neti, J. Luettin & I. Matthews. Issues in Visual and Audio-Visual Speech Processing: Audio-Visual Automatic Speech Recognition: An Overview. MIT Press, 2004.

[Pusateri-02] E. J. Pusateri & T. J. Hazen. *Rapid speaker adaptation using speaker clustering.* Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002), str. 165–168, Denver, Kolorado, ZDA, september 2002.

[Rabiner-89] L. R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, vol. 77, št. 2, str. 257–286, februar 1989.

[Ramos-Castro-05] D. Ramos-Castro, D. Garcia-Romero, I. Lopez-Moreno & J. Gonzalez-Rodriguez. *Speaker verification using fast adaptive TNORM based on Kullback–Leibler divergence.* Third COST 275 Workshop: Biometrics on the Internet, University of Hertfordshire, Velika Britanija, oktober 2005.

[Reynolds-95] D. A. Reynolds. *Speaker identification and verification using Gaussian mixture speaker models.* Speech Communication, vol. 17, št. 1–2, str. 91–108, avgust 1995.

[Reynolds-03a] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones & B. Xiang. *The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition.* IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), vol. 4, str. 784–787, Hong Kong, Kitajska, april 2003.

[Reynolds-03b] D. A. Reynolds, J. P. Campbell, W. M. Campbell, R. B. Dunn, T. P. Gleason, D. A. Jones, T. F. Quatieri, C. B. Quillen, D. E. Sturim & P. A. Torres-Carrasquillo. *Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition.* Proceedings of the Workshop on Multimodal User Authentication, str. 223–229, Santa Barbara, Kalifornija, ZDA, december 2003.

[Reynolds-05] D. A. Reynolds & P. A. Torres-Carrasquillo. *Approaches and Applications of Audio Diarization.* Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005), Philadelphia, ZDA, marec 2005.

[Roeder-97] K. Roeder & L. Wasserman. *Practical Bayesian density estimation using mixtures of normals.* Journal of the American Statistical Association, vol. 92, str. 894–902, 1997.

[Saggion-04] H. Saggion, H. Cunningham, K. Bontcheva, D. majnard, O. Hamza & Y. Wilks. *Multimedia indexing through multisource and multi-language information extraction: the MUMIS project.* Data & Knowledge Engineering, vol. 48, št. 2, str. 247–264, februar 2004.

[Samouelian-98] A. Samouelian, J. Robert-Ribes & M. Plumpe. *Speech, silence, music and noise classification of TV broadcast material.* International Conference on Spoken Language Processing (ICSLP98), vol. 3, str. 1099–1102, Sydney, Avstralija, 30. november – 4. december 1998.

[Saraclar-02] M. Saraclar, M. Riley, E. Bocchieri & V. Goffin. *Towards Automatic Closed Captioning : Low Latency Real Time Broadcast News Transcription.* Proceedings of the International Conference on Speech and Language Processing (ICSLP '02), Denver, Kolorado, ZDA, september 2002.

[Saunders-96] J. Saunders. *Real-time discrimination of broadcast speech/music.* International Conference on Acoustics, Speech, and Signal Processing (ICASSP 96), vol. 2, str. 993–996, Atlanta, ZDA, 1996.

[Scheirer-97] E. Scheirer & M. Slaney. *Construction and evaluation of a robust multifeature speech/music discriminator.* Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing, ICASSP 1997, vol. 2, str. 1331–1334, München, Nemčija, april 1997.

[Scwartz-76] G. Scwartz. *Estimating the dimension of a model.* Annals of Statistics, vol. 6, str. 461–464, 1976.

[Shafran-03] I. Shafran & R. Rose. *Robust speech detection and segmentation for real-time ASR applications.* Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing, ICASSP 2003, Hong Kong, Kitajska, april 2003.

[Shriberg-00] E. Shriberg, A. Stolcke & G. Tür D. Hakkani-Tür. *Prosody-based automatic segmentation of speech into sentences and topics.* Speech Communication, vol. 32, št. 1-2, str. 127–154, september 2000.

[Shriberg-05] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman & A. Stolcke. *Modeling prosodic feature sequences for speaker recognition.* Speech Communication, vol. 46, št. 3–4, str. 455–472, julij 2005.

[Siegler-97] M. Siegler, U. Jain, B. Raj & R. Stern. *Automatic Segmentation, Classification and Clustering of Broadcast News Data,.* Proceedings of the DARPA Speech Recognition Workshop, str. 97–99, Chantilly, VA, ZDA, 1997.

[Sinha-05] R. Sinha, S. E. Tranter, M. J. F. Gales & P. C. Woodland. *The Cambridge University marec 2005 Speaker Diarisation System.* Proceedings of Interspeech 2005 - Eurospeech, str. 2437–2440, Lizbona, Portugalska, september 2005.

[Siohan-01] O. Siohan, A. Ando, M. Afify, H. Jiang, C.-H. Lee, Q. Li, F. Liu, K. Onoe, F. K. Soong & Q. Zhou. *A real-time Japanese broadcast news closed-captioning system.* Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001), str. 495–498, Aalborg, Danska, 2001.

[Solomonoff-98] A. Solomonoff, A. Mielke, M. Schmidt & H. Gish. *Clustering speakers by their voices.* Proceedings of the ICASSP 98, str. 757–760, Seattle, Washington, ZDA, maj 1998.

[Talkin-95] D. Talkin. *Chapter 14: A robust algorithm for pitch tracking (RAPT).* W. B. Kleijn & K. K. Paliwal, (ur.), Speech coding and synthesis, str. 495–518. Elsevier Science, 1995.

[Theodoridis-03] S. Theodoridis & K. Koutroumbas. Pattern Recognition (2nd edition). Academic Press, januar 2003.

[Tibshirani-00] R. Tibshirani, G. Walther & T. Hastie. *Estimating the Number of Clusters in a Dataset via the Gap Statistic.* Tehnično poročilo, Department of Biostatistics, Stanford University, marec 2000.

[Tritschler-99] A. Tritschler & R. Gopinath. *Improved speaker segmentation and segments clustering using the Bayesian information criterion.* Proceedings of EUROSPEECH 99, vol. 2, str. 679–682, Budimpešta, Madžarska, september 1999.

[Vandecatseye-03] A. Vandecatseye & J. P. Martens. *A fast, accurate and stream-based speaker segmentation and clustering algorithm.* Proceedings of the Eurospeech 2003, vol. 2, str. 941–944, Ženeva, Švica, 2003.

[Vandecatseye-04] A. Vandecatseye, J. P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza,

P. David, M. Pleva, A. Cizmar, H. Papageorgiou & C. Alexandris. *The COST278 pan-European Broadcast News Database*. Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004), str. 873–876, Lizbona, Portugalska, maj 2004.

[Viswanathan-00] M. Viswanathan, H. Beigi, S. Dharanipragada, F. Maali & L. Tritschler. *Multimedia document retrieval using speech and speaker recognition*. International Journal on Document Analysis and Recognition, vol. 2, št. 4, str. 147–162, junij 2000.

[Wactlar-99] H. Wactlar, M. Christel, Y. Gong & A. Hauptmann. *Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library*. IEEE Computer, vol. 32, št. 2, str. 66–73, 1999.

[Waibel-04] A. Waibel, H. Steusloff, R. Stiefelhagen & the CHIL Project Consortium. *Chil: Computers In The Human Interaction Loop*. 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), vol. 1, Lizbona, Portugalska, april 2004.

[Walls-99] F. Walls, H. Jin, S. Sista & S. Schwartz. *Topic Detection in Broadcast News*. Proceedings of the DARPA Broadcast News Workshop, str. 193–198, Herndon VA, ZDA, februar 1999.

[Wayne-00] C. Wayne. *Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation*. Proceedings of Language Resources and Evaluation Conference (LREC) 2000, str. 1487–1494, Atene, Grčija, maj–junij 2000.

[Weintraub-93] M. Weintraub. *Keyword-spotting using SRI's DECIPHER large-vocabulary speech recognition system*. Proceedings of the ICASSP 93, vol. 2, str. 463–466, Minneapolis, ZDA, 1993.

[Williams-99] G. Williams & D. P. W. Ellis. *Speech/music discrimination based on posterior probabilities*. Proceedings of EUROSPEECH 99, str. 687–690, Budimpešta, Madžarska, september 1999.

[WinCAPS-06] Sysmedia. *Professional Subtitling Software*, 2006. `http://www.sysmedia.com/sysmedia_subtitling/subtitling_overview.asp`.

[Woodland-02] P. C. Woodland. *The development of the HTK Broadcast News transcription system: An overview*. Speech Communication, vol. 37, št. 1-2, str. 47–67, maj 2002.

[Yamron-00] J. P. Yamron, L. Gillick, S. Knecht, S. Lowe & P. van Mulbregt. *Statistical models for tracking and detection.* In Working Notes fo the the Third TDT-3 Workshop, 2000.

[Young-04] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Vatchev & P. Woodland. The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department, Cambridge, Velika Britanija, 2004.

[Zdansky-05] J. Zdansky & J. Nouza. *Detection of Acoustic Change-Points in Audio Records via Global BIC Maximization and Dynamic Programming.* Proceedings of Interspeech 2005 - Eurospeech, Lizbona, Portugalska, september 2005.

[Zhang-02] Z. Zhang, S. Furui & K. Ohtsuki. *On-line incremental speaker adaptation for broadcast news transcription.* Speech Communication, vol. 37, št. 3-4, str. 271–281, 2002.

[Zhou-00] B. Zhou & J. H. L. Hansen. *Unsupervised audio stream segmentation and clustering via the Bayesian information criterion.* Proceedings of the International Conference on Spoken Language Processing, ICSLP 2000, str. 714–717, Peking, Kitajska, oktober 2000.

[Zhu-05] X. Zhu, C. Barras, S. Meignier & J.-L. Gauvain. *Combining Speaker Identification and BIC for Speaker Diarization.* Proceedings of Interspeech 2005 - Eurospeech, str. 2437–2440, Lizbona, Portugalska, september 2005.

[Zissman-96] M. A. Zissman. *Comparison on four approaches to automatic language identification for telephone speech.* IEEE Transactions on Speech and Audio Processing, vol. 4, št. 1, str. 31–44, januar 1996.

[Žgank-04] A. Žgank, T. Rotovnik, D. Verdonik & Z. Kačič. *Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora.* J. G. Žganec T. Erjavec, (ur.), Jezikovne tehnologije : zbornik mednarodne multi-konference Informacijska družba IS 2004, str. 94–97, Ljubljana, Slovenija, oktober 2004.

[Žibert-00] J. Žibert, F. Mihelič & S. Dobrišek. *Avtomatično podnaslavljanje vremenskih napovedi.* Zbornik devete Elektrotehniške in računalniške konference ERK 2000, str. 165–168. IEEE Region 8, Slovenska sekcija IEEE, september 2000. Portorož, Slovenija.

[Žibert-04] J. Žibert & F. Mihelič. *Development, evaluation and automatic segmentation of Slovenian broadcast news speech database.* J. G. Žganec T. Erjavec, (ur.), Jezikovne tehnologije :

zbornik mednarodne multi-konference Informacijska družba IS 2004, str. 72–78, Ljubljana, Slovenija, oktober 2004.

[Žibert-05] J. Žibert, F. Mihelič, J.-P. Martens, H. Meinedo, J. Neto, L. Docio, C. Garcia-Mateo, P. David, J. Zdansky, M. Pleva, A. Cizmar, A. Žgank, Z. Kačič, C. Teleki & K. Vicsi. *The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results.* Proceedings of Interspeech 2005 - Eurospeech, str. 629–632, Lizbona, Portugalska, september 2005.

[Žibert-06] J. Žibert, N. Pavešić & F. Mihelič. *Speech/Non-Speech Segmentation Based on Phoneme Recognition Features.* EURASIP Journal on Applied Signal Processing, vol. 2006, str. 1–13, 2006. Article ID 90495.

Slovar izrazov

agglomerative clustering	rojenje z združevanjem
audio indexing	indeksacija zvočnih posnetkov
automatic audio indexing	samodejna indeksacija zvočnih posnetkov
automatic broadcast news transcription	samodejno označevanje informativnih oddaj
automatic summarization	samodejna izdelava povzetkov
background change detection	(pri segmentaciji) točka spremembe akustičnega ozadja
background model	model ozadja
Bayesian information criterion, BIC	Bayesov informacijski kriterij, kriterij BIC
bottom-up clustering	rojenje z združevanjem od spodaj navzgor
broadcast news speech database	govorna podatkovna zbirka informativnih oddaj
cepstral mean and variance normalization, CMVN	normalizacija z izničevanjem skupnega povprečja in variance kepstralnih značilk
cluster purity	mera čistosti roja; mera, kako dobro z enim rojem segmentov opišemo določenega govorca
confidence measure	mera zaupanja v oceno
consonant–vowel, CV	par soglasnik–samoglasnik
content-based audio retrieval	pridobivanje vsebinskih informacij iz zvočnih zapisov
cross log-likelihood ratio, CLR	navzkrižni kriterij razmerij logaritmov verjetnostnih ocen
data stream	(pri fuziji) tok podatkov
diarisation error rate, DER	mera napake med ujemanjem referenčnih ter samodejno pridobljenih in označenih segmentov, mera DER
dynamism	dinamizem
entropy	entropija
episode	(pri označevanju) oddaja

expiration event	(pri označevanju) področje izdiha
F–measure	mera F pri ocenjevanju segmentacije
false alarm rate, FAR	delež napačnih zaznavanj
false alarm	napačno zaznavanje; (pri meri DER) napaka zaznavanja
false rejection rate, FRR	delež napačnih zavrnitev
false rejection	napačna zavrnitev
feature warping, FW	postopek prileganja značilk k normalnim porazdelitvam
focus conditions, F–conditions	(pri označevanju) F–stanja
forced alignment	postopki vsiljenega prileganja
frame skip	premik po času za naslednji izračun
Gaussian mixture model, GMM	model kombinacije Gaussovih porazdelitev, GMM model
hidden Markov model, HMM	prikrit Markovov model, HMM model
(high) segment purity	(visoka stopnja) čistosti segmentov
independent and identically distributed, IID	enako porazdeljene in med seboj neodvisne (naključne spremenljivke)
inspiration event	(pri označevanju) področje vdiha
keyword indexing	indeksacija podatkov po ključnih besedah
keyword spotting	detekcija ključnih besed
large vocabulary continuous speech recognition system, LVCSRs	sistem za razpoznavanje tekočega govora velikega števila različnih besed
likelihood	verjetnostna ocena
linear prediction coefficients, LPC	koeficienti linearne predikcije
link detection	(pri obdelavi vsebin inf. oddaj) povezovanje področij sorodnih vsebin
log-likelihood ratio, LLR	razmerje logaritmov verjetnostnih ocen
log-likelihood, LLH	logaritem verjetnostne ocene
maximum a posteriori adaptation, MAP adaptation	postopek prilagajanja modelov z večanjem aposteriornih verjetnosti, MAP adaptacija
maximum likelihood estimate	izračun največje verjetnostne ocene
mel–frequency cepstral coefficients, MFCC	koeficienti melodičnega kepstra, MFCC koeficienti

miss error	(pri meri DER) napaka zgrešitve
multi-lingual audio indexing	večjezikovna indeksacija zvočnih posnetkov
pitch, f_0	višina osnovnega tona v govornem signalu
precision, PRC	natančnost pri ocenjevanju segmentacije
recall, RCL	priklic pri ocenjevanju segmentacije
rich transcription	(pri označevanju) dodatna informacija poleg tekstovnega prepisa zvočnih posnetkov
score normalization	(pri fuziji) normalizacija ocen odločitev
score-based fusion	fuzija na podlagi združevanja ocen odločitev
section	(pri označevanju) sekcija
segments tying	vezava segmentov pri samodejni indeksaciji zvočnih posnetkov
segment	segment, odsek
speaker adaptation	postopek prilagajanja modelov posameznim govorcem
speaker change detection	(pri segmentaciji) točka zamenjave govorcev
speaker clustering	rojenje segmentov po govorcih
speaker diarisation	segmentacija in združevanje segmentov po govorcih
speaker error	(pri meri DER) napaka ujemanja med roji in referenčnimi govorci
speaker purity	mera, kako dobro je en govorec opisan z določenim rojem segmentov
speaker segments tying	vezava segmentov po govorcih
speaker turn	(pri označevanju) odsek govorca
speaker–adapted training, SAT	postopek prilagajanja splošnih modelov govora glede na govorca
speaker-based index	indeks govorcev
spectral centroid	spektralno središče
spectral flux	spektralni tok
speech/non–speech segmentation	segmentacija zvočnih posnetkov na govorne in ne–govorne odseke
story segmentation	segmentacija na področja zaključenih vsebin novic

support vector machine, SVM	metoda podpornih vektorjev, SVM
target speaker model	model iskanega govorca
text and audio/video data alignment	časovna poravnava avdio/video posnetkov s tekstovno predlogo
topic detection and tracking	detekcija in sledenje vsebinam informativnih oddaj
topic detection	detekcija vsebin informativnih oddaj
universal background model, UBM	splošen model govora
vector quantisation distortion criterion	kriterij popačenja vektorske kvantizacije
voiced–unvoiced, VU	par zveneči–nezveneči glas
word error rate, WER	napaka razpoznavanja govora
zero-crossing rate, ZCR	(pri obdelavi signalov) število prehodov signala skozi nič

Izvirni prispevki k znanosti

Disertacija vsebuje naslednje pomembnejše izvirne prispevke k znanosti:

- **Postopek pridobivanja fonetičnih značilk za ugotavljanje govornih odsekov v zvočnih posnetkih**.

  V razdelku 3.2 disertacije smo predlagali nov postopek pridobivanja fonetičnih značilk na podlagi zaporedij razpoznanih osnovnih govornih enot, ki smo jih pridobili s preprostimi sistemi razpoznavanja glasov iz zvočnih posnetkov. Izpeljali smo štiri značilke, ki so temeljile na trajanju in spremembah dveh skupin govornih enot: parov samoglasnik–soglasnik (CVS značilke) ter zvenečih in nezvenečih glasov (VUS značilke).

  Izvedbi fonetičnih značilk smo prilagodili tudi postopek segmentacije zvočnih posnetkov na govorne in ne–govorne odseke, ki smo ga predlagali v razdelku 3.3. Medtem ko se v osnovnem postopku izvaja segmentacija in razvrščanje segmentov na govor in ne–govor hkrati z vključevanjem modelov kombinacije Gaussovih porazdelitev (GMM) v mreže prikritih Markovovih modelov (HMM), smo v predlaganem postopku najprej izvajali segmentacijo glede na akustične spremembe v zvočnih posnetkih in nato razvrščanje segmentov na govor in ne–govor z uporabo fonetičnih značilk.

  Ukvarjali smo se tudi z združevanjem osnovnih akustičnih in predlaganih fonetičnih značilk. Tako je bil v razdelku 3.4.3 predstavljen postopek segmentacije s fuzijo, s katerim smo v različnih preizkusih dosegli najboljše rezultate detekcije govora.

- **Postopek segmentacije z združevanjem različnih akustičnih predstavitev zvočnih posnetkov in z relativno določenim pragom odločitvene kriterijske funkcije**.

  Pri samodejni segmentaciji zvočnih posnetkov glede na zamenjave govorcev in spremembe v akustičnem ozadju smo se v poglavju 4 osredotočili predvsem na izvedbo postopkov segmentacije, ki bi bili čim manj odvisni od izbire odprtih parametrov postopkov in s tem manj občutljivi na spremembe akustičnih razmer v zvočnih posnetkih. Tako smo v razdelku 4.4.1 predlagali postopek segmentacije z relativno določenim pragom odločitve, kjer smo združili dva obstoječa postopka segmentacije: standardni postopek z Bayesovim informacijskim kriterijem (kriterij BIC) in postopek DISTBIC. S postopkom DISTBIC smo v prvi fazi ocenili možne vrednosti kriterija BIC in s tem prag odločitve za meje, s standardnim postopkom v drugi fazi pa smo določili meje med segmenti. Vhodni parameter v postopek tako ni bil več absolutni prag odločitve, ampak relativno določeni prag

glede na ocenjene vrednosti kriterija BIC, ki so se spreminjale glede na akustične lastnosti zvočnih posnetkov.

Možnost ocenjevanja vrednosti kriterija BIC smo izrabili tudi v drugem predlaganem postopku segmentacije (razdelek 4.4.2), ki je temeljil na združevanju različnih akustičnih predstavitev zvočnih signalov. Na podlagi ocen kriterijev BIC različnih predstavitev smo lahko izvajali normalizacijo ocen posameznih kriterijev in s tem združevanje ocen s postopki fuzije. Na ta način smo tako z združevanjem ločenih predstavitev lahko bolj zanesljivo ocenjevali krajše odseke v zvočnih posnetkih, ki jih v primeru skupnih predstavitev pri standardnem postopku segmentacije slabo modeliramo.

V preizkusih na razvojnih posnetkih informativnih oddaj (razdelek 4.5.3) smo lahko potrdili večjo zanesljivost in neobčutljivost predlaganih postopkov v primeru različnih (neoptimalnih) izbir pragov odločitve in ostalih odprtih parametrov. Prav tako smo s predlaganimi postopki dosegli tudi boljše rezultate na obsežnih testnih zbirkah posnetkov (razdelek 4.5.4) v primerjavi z obstoječimi postopki ob optimalni izbiri parametrov.

- **Postopek združevanja segmentov po govorcih z uporabo akustične in prozodične informacije**.

  V razdelku 5.4 smo predstavili postopek rojenja segmentov po govorcih z združevanjem akustične in prozodične informacije. Osnovna ideja je bila, da bi osnovnim akustičnim predstavitvam segmentov dodali še prozodično informacijo, ki bi bila primerna za združevanje segmentov po govorcih. V ta namen smo izpeljali 10 prozodičnih značilk, ki smo jih pridobivali iz energije signala, ocene osnovnega tona v signalu in na podlagi razpoznanih osnovnih glasovnih enot v signalu. Na ta način smo vpeljali informacijo višjega reda v postopke rojenja, s čimer smo želeli izboljšati združevanje segmentov v primeru, ko se samo na podlagi akustične informacije ali pa zaradi slabih akustičnih razmer ne bi znali pravilno odločati za združevanje med posameznimi roji segmentov. Z vpeljavo dodatne prozodične informacije osnovnim akustičnim značilkam smo morali prilagoditi tudi osnovni postopek rojenja, da bi lahko potekalo združevanje segmentov na podlagi kombinacije obeh predstavitev.

  V vseh preizkusih smo s predlaganim postopkom rojenja segmentov z združevanjem akustične in prozodične informacije izboljšali rezultate razvrščanja segmentov po govorcih v primerjavi z obstoječimi postopki, kjer se izvaja rojenje segmentov samo z uporabo akustične informacije (razdelka 5.5.3 in 5.5.4). Skupni končni rezultati razvrščanja pa so še dodatno potrdili zanesljivost in neobčutljivost delovanja predlaganega postopka v primeru različnih akustičnih lastnosti zvočnih posnetkov in različnih kriterijev zaustavitve postopkov združevanja segmentov.

- **Postopek razvrščanja segmentov po govorcih z vpeljavo nove mere združevanja segmentov**.

  V razdelku 5.3.2 je opisana izvedba postopka rojenja z združevanjem segmentov po govorcih z uporabo metod razpoznavanja govorcev. Tu smo osnovne segmente govora predstavili z GMM modeli, ki smo jih izpeljali iz splošnih modelov govora

(UBM) ob uporabi MAP adaptacije. Pri tem postopku smo se ukvarjali predvsem z različnimi kriteriji združevanja tako predstavljenih segmentov in predlagali novo mero, ki je temeljila na kriteriju BIC (razdelek 5.3.2.1).

V vseh preizkusih rojenja segmentov po govorcih smo s predlagano mero združevanja znatno presegli rezultate razvrščanja v primerjavi s postopkom, kjer smo za mero združevanja uporabljali obstoječi kriterij (razdelka 5.5.3 in 5.5.4). V primeru preizkusov rojenja ročno določenih segmentov (razdelek 5.5.3) pa smo celo dosegli najboljše skupne rezultate razvrščanja segmentov po govorcih.

- **Kriterija za zaustavitev postopkov združevanja segmentov, ki temeljita na Bayesovem informacijskem kriteriju in relativni oceni napake združevanja za dva različna postopka rojenja**.

  Pri raziskovanju kriterijev zaustavitve rojenja v razdelku 5.6 smo želeli poiskati takšne pristope, s katerimi bi se v postopkih rojenja ustavili v tistih točkah, kjer bi dosegli najmanjšo možno napako ujemanja med dejanskimi segmenti govorcev in združenimi segmenti, ki jih pridobimo pri rojenju. Izpeljali smo dva kriterija. Prvi je temeljil na skupnem kriteriju BIC (razdelek 5.6.2.1) in je primeren v postopkih rojenja, kjer se pri združevanju prav tako uporablja kriterij BIC. Drugi kriterij zaustavitve je temeljil na relativni oceni mere DER (razdelek 5.6.2.2). Pri tem kriteriju smo potrebovali dva različna postopka združevanja, kjer smo na podlagi primerjave napak enega postopka z drugim ocenjevali možne točke zaustavitve postopkov rojenja.

  V preizkusih iskanja končnega števila rojev v kombinaciji z različnimi postopki rojenja (razdelek 5.7) smo se z obema predlaganima postopkoma zelo približali referenčnim rezultatom rojenja, kjer smo zaustavitev postopkov definirali z dejanskim številom govorcev v posnetkih. V primerjavi z osnovnim kriterijem zaustavitve pa smo v povprečju končne rezultate znatno presegli.