Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Using Topic Cues for Speaker Recognition in Broadcast Multimedia Archives

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der technischen Wissenschaften

eingereicht von

Doris Baum

Matrikelnummer 0426038

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: a.o. Prof. Dr. Andreas Rauber

Diese Dissertation haben begutachtet:

(a.o. Prof. Dr. Andreas Rauber)

(apl. Prof. Dr. Frank Kurth, Universität Bonn)

Wien, 10.05.2013

(Doris Baum)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Using Topic Cues for Speaker Recognition in Broadcast Multimedia Archives

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der technischen Wissenschaften

by

Doris Baum
Registration Number 0426038

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: a.o. Prof. Dr. Andreas Rauber

The dissertation has been reviewed by:

(a.o. Prof. Dr. Andreas Rauber)

(apl. Prof. Dr. Frank Kurth, Universität Bonn)

Wien, 10.05.2013

(Doris Baum)

Erklärung zur Verfassung der Arbeit

Doris Baum
Streffleurgasse 6/1, 1200 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasserin)

Acknowledgements

To my dad, for raising and encouraging my interest in science.

I would like to thank Professor Andreas Rauber for supervising this thesis, for offering advice and guidance and at the same time giving me intellectual freedom.
Special thanks go to Professor Frank Kurth for helpful discussion and accepting to act as second examiner.
Also, thanks go to my kind and helpful former colleagues at Fraunhofer IAIS, with special thanks to Respected Doctor Rolf Bardeli for taking me on as his acolyte and to Jochen Schwenninger and Mark Asbach for fruitful discussions and open ears.
Thanks to the Speech Processing Group at the Brno University of Technology for welcoming me on their team for the NIST 2010 evaluations.
Thanks to my mum for loving support.
Thanks to Christian for having my back in times of trouble.

Part of this study was funded by the German Federal Ministry of Economics and Technology through the CONTENTUS scenario of the THESEUS project.

Abstract

Annotating and indexing large amounts of broadcast multimedia material in order to make it easily accessible to archive users is currently a big challenge for broadcasters and archives. Because of the sheer amount of data, purely manual annotation is often infeasible, and automatic or semi-automatic annotation methods are required to alleviate the problem. For audio data it is not only important to know what was said (automatic speech recognition) but also to reliably determine who said it, which requires effective speaker recognition methods.

Current speaker recognition research focusses on telephone data applications and on features which capture the voice of the speaker. However, especially in the broadcast domain, other cues are linked with the speakers and can be used to identify them, such as the topic of an utterance: Speakers often appear in programmes on topics for which they are considered experts. For example, a tennis player is more likely to talk about tennis and to appear in a report about sports than a politician. Vice versa, one can assume that a report about sports is more likely to contain a tennis player than a politician.

Using topic cues to automatically identify speakers seems to have been largely disregarded in the speaker recognition literature. Thus, the goal of this thesis is to explore the idea of topic-based speaker recognition for broadcast data. It tries to answer the question how spoken utterances' topic information can be automatically captured with little manual effort, and how it can be used to learn speakers' topic preferences. Two approaches which automatically learn the speakers' topic preferences from audio training data are presented, one based on implicit and one based on explicit topic representation. Automatic speech recognition is used to produce word transcripts for the audio material, as the topic information is carried mostly by the words in an utterance. In the first approach, the speakers' topic preferences are implicitly represented by learning their idiosyncratic words, which will consist mostly of topic marker words for their preferred topics. In the second approach, topics are explicitly modelled using an unsupervised probabilistic topic modelling algorithm (Latent Dirichlet Allocation) which automatically identifies prevalent topics and their marker words without need for manually labelled topic training data. With the explicitly trained topics, the utterances' word transcripts can be converted into topic probability vectors which are then used to model and recognise speakers.

Both approaches are evaluated to see how speaker identification based on topic cues performs. As no big speaker recognition evaluation benchmark containing topic preferences was available for the evaluation, a new corpus based on recordings from the German parliament was created. The evaluation compares topic-based with traditional voice-based and idiolectal speaker recognition systems. Also, topic, voice, and idiolectal systems are fused to see how well topic combines with other cues.

As speaker interests change over time, consideration is also given to the effect of changing topic preferences over a larger timespan.

Kurzfassung

Das Annotieren und Indexieren großer Mengen von Rundfunksendungen, um sie für Archivbenutzer einfach verfügbar zu machen, ist zur Zeit eine große Herausforderung für Sendeanstalten und Archive. Eine rein manuelle Annotation ist aufgrund der großen Datenmengen oft nicht machbar, kann aber durch den Einsatz von automatischen oder semi-automatischen Methoden ersetzt werden. Bei der Analyse von Audiodaten ist es nicht nur wichtig herauszufinden, was gesagt wurde (Automatische Spracherkennung), sondern auch, wer der Sprecher ist. Dies erfordert effektive Methoden zur automatischen Sprechererkennung.

Die derzeitige wissenschaftliche Forschung zur Sprechererkennung ist stark fokussiert auf die Analyse von Telefondaten und die Verwendung von Merkmalen, die die Stimme des Sprechers abbilden. Gerade im Rundfunkbereich gibt es aber noch andere sprecherspezifische Merkmale, die genutzt werden können, um Sprecher zu identifizieren — beispielsweise das Thema einer Äußerung. Häufig treten Sprecher in Sendungen auf, für deren Themen sie als Experten gelten. Ein Tennisspieler wird zum Beispiel mit höherer Wahrscheinlichkeit über Tennis reden und in Sportsendungen auftreten als ein Politiker. Umgekehrt kann man annehmen, dass ein Bericht über Sport mit größerer Wahrscheinlichkeit die Äußerungen eines Tennisspielers enthält als die eines Politikers.

In der bisherigen Fachliteratur zur Sprechererkennung scheint die Verwendung von Themenmerkmalen zur automatischen Identifizierung von Sprechern keine große Beachtung gefunden zu haben. Deshalb ist es das Ziel dieser Arbeit, das Konzept der themenbasierten Sprechererkennung für den Rundfunkbereich zu untersuchen. Es soll die Frage beantwortet werden, wie die Themeninformation einer Äußerung mit möglichst wenig manuellem Aufwand erfasst werden kann, und wie damit die spezifischen Themen eines Sprechers automatisch gelernt werden können. Es werden zwei Ansätze vorgestellt, die die Lieblingsthemen der Sprecher automatisch aus Audio-Trainingsmaterial lernen, wobei einer auf impliziter und der andere auf expliziter Repräsentation der Themen basiert. Ein automatisches Spracherkennungssystem wird eingesetzt, um für Audiomaterial Worttranskripte zu erzeugen, da die Themeninformation hauptsächlich in den Worten einer Äußerung enthalten ist. Im ersten Ansatz werden die Themenvorlieben der Sprecher implizit modelliert, indem ihre idiosynkratischen Lieblingswörter ge-

lernt werden — diese sind üblicherweise hauptsächlich Themenschlüsselwörter der bevorzugten Themen. Im zweiten Ansatz werden die Themen explizit modelliert: Mit einem unüberwachten probabilistischen Themenmodellierungs-Algorithmus (Latent Dirichlet Allocation) werden automatisch wichtige Themen und deren Schlüsselwörter identifiziert, ohne dass dazu manuell annotierte Themen-Trainingsdaten notwendig sind. Mit dem explizit trainierten Themenmodell ist es dann möglich, Worttranskripte von Äußerungen in Themen-Wahrscheinlichkeitsvektoren umzuwandeln, die verwendet werden, um Sprecher zu modellieren und zu erkennen.

Beide Ansätze werden evaluiert um zu bewerten wie gut themenbasierte Sprechererkennung funktioniert. Da kein großes Korpus zur Evaluierung von Sprechererkennungssystemen verfügbar war, das Themenvorlieben der Sprecher enthielt, wurde ein neues Korpus aus Aufnahmen aus dem deutschen Bundestag geschaffen. Die resultierende Evaluierung vergleicht die themenbasierte Sprechererkennung mit einem traditionellen stimmbasierten und einem idiolektbasierten System. Außerdem werden die themen-, stimm-, und idiolektbasierten Systeme kombiniert, um zu sehen, wie gut Themeninformation die anderen Merkmale ergänzt. Da sich die Interessen der Sprecher mit der Zeit verschieben, wird auch der Effekt sich verändernder Themenvorlieben untersucht.

Contents

Contents **xiii**

1 Introduction **1**

1.1 Research Questions . . . 3

1.2 Outline . . . 4

2 Related Work **7**

2.1 Performance Estimation . . . 9

2.1.1 Performance Measures . . . 9

2.1.2 NIST Speaker Recognition Evaluation Benchmarks . . . 11

2.1.3 Automatic Speech Recognition Benchmarks . . . 12

2.2 Cues and Features for Speaker Recognition . . . 13

2.2.1 Voice-based / Spectral Speaker Recognition . . . 14

2.2.2 Prosodic Speaker Recognition . . . 15

2.2.3 Pronunciation-based / Phonetic Speaker Recognition . . . 16

2.2.4 Idiolectal / Word-based Speaker Recognition . . . 19

2.2.5 Naming unknown speakers . . . 20

2.2.6 Authorship Attribution for Written Texts . . . 20

2.3 Speaker Modelling and Classification . . . 22

2.3.1 Log-likelihood Ratios . . . 22

2.3.2 Gaussian Mixture Models . . . 23

2.3.3 Support Vector Machines . . . 26

2.3.4 Joint Factor Analysis . . . 27

2.4 Normalisation and Fusion . . . 29

2.4.1 Feature Normalisation . . . 29

2.4.2 Score normalisation . . . 30

2.4.3 Fusion . . . 30
2.5 Summary . . . 31

3 Topic-based Speaker Recognition **33**
3.1 Motivation . . . 33
3.2 Constraints for Topic Information Extraction . . . 34
3.3 Utterance Pre-processing . . . 35
3.4 Keyword-based Speaker Recognition . . . 37
3.4.1 Log-likelihood Ratio Scoring . . . 37
3.4.2 Support Vector Machine Classifier . . . 40
3.5 Topic-model-based Speaker Recognition . . . 43
3.5.1 Topic Derivation . . . 44
3.5.2 Feature Extraction . . . 49
3.5.3 Restrictions on Speaker Modelling with Topic Feature Vectors . . . 50
3.5.4 Modelling and Scoring with Speaker Prototype Vectors . . . 50
3.5.5 Modelling and Scoring with Lazy Learning . . . 51
3.5.6 Modelling and Scoring with Support Vector Machines . . . 52
3.6 Summary . . . 53

4 The Bundestag Corpus **56**
4.1 Corpus Requirements . . . 57
4.2 Corpus Preparation . . . 58
4.2.1 Consistency Checking . . . 58
4.2.2 Evaluation Protocol Design . . . 59
4.3 Corpus Characteristics . . . 61
4.4 Text Data from the Bundestag Domain . . . 62
4.4.1 Speaker Recognition Background Set . . . 64
4.4.2 Speaker Preference Shift Evaluation Corpus . . . 64
4.5 Summary . . . 65

5 Experimental Evaluation **67**
5.1 System Setup . . . 67
5.1.1 Utterance Pre-processing . . . 67
5.1.2 Keyword-based System Setup . . . 69
5.1.3 Topic-model-based System Setup . . . 69

5.2 Baseline Systems 70
5.2.1 Spectral Speaker Recognition 71
5.2.2 Idiolectal Speaker Recognition 71
5.3 System Fusion 72
5.4 Significance Testing for Evaluation Results 73
5.5 Calibration 74
5.5.1 Modelling for Keyword-based Speaker Recognition 74
5.5.2 Influence of Pre-processing on Keyword-based Speaker Recognition 74
5.5.3 Influence of Term Weighting on Keyword-based Speaker Recognition 76
5.5.4 Optimal Number of Topics for Topic-model-based Speaker Recognition 77
5.5.5 Distance Measure for Topic-model-based Speaker Recognition 78
5.5.6 Modelling for Topic-model-based Speaker Recognition 80
5.5.7 Final System Configuration for the Topic-based Systems 81
5.6 Evaluation Results 82
5.6.1 Subjective Evaluation of Speaker-Topic Assignment 82
5.6.2 Performance Comparison of the Different Speaker Cues 84
5.6.3 Influence of the Length of the Test Utterance on Topic-based Speaker Recognition 86
5.6.4 Fusion of Topic-based Systems with Baseline Systems 89
5.6.5 Analysis of Remaining Misclassifications 94
5.7 Influence of Speaker Interest Changes over Time 97
5.8 Summary 101

6 Conclusion and Future Work **104**

Bibliography **108**

List of Figures **122**

List of Tables **124**

Chapter 1

Introduction

As the amount of produced multimedia content increases every day for TV and radio broadcasters, and as large data digitisation projects are currently under way at a number of libraries and archives[1], manual analysis and annotation of the material becomes an infeasible task. However, metadata describing the structure and content of the multimedia data is essential for efficiently searching and retrieving it from the ever-growing archives, enabling users to do research and broadcasters to re-use existing material. A lot of content is produced by broadcasters, shown on TV once, and then stored and forgotten because the archives lack manpower to annotate it. Automatic or semi-automatic annotation can alleviate this problem and even open up new ways to use the data which were previously unavailable. Using automatic methods, the forgotten material can be indexed and found by users. Journalists may research what was previously reported on an issue more easily and are enabled to re-use recorded footage. Web users can search for specific content in the broadcasters' online archives, and the found content can be accessed and browsed more easily if its structure is known.

For spoken content, it is important to know not only what is being said but also who says it, as this allows searching not only for keywords or topics but also for the people talking about specific subjects. During an electoral campaign, voters might for example search for statements about a political issue from the competing politicians in order to compare them. In 2009, a multimedia search engine was built by German broadcaster ARD based on automatic analysis, enabling users to search by speaker for topics and keywords in a large set of broadcasts involving the leaders of all major parties[2]. This would have been extremely costly to realise based only on manual annotations of the data.

To automatically annotate large amounts of material with speaker information requires a good speaker recognition system. A low recognition error rate is important particularly when lots of data is processed,

[1] For example `http://www.europeana.eu` and `http://www.deutsche-digitale-bibliothek.de/` last accessed: 2013-02-25

[2] `http://www.ard.de/-/id=1208788/7xuyb4/index.html`, last accessed: 2013-02-25

as the absolute amount of errors obviously increases with the number of recordings annotated and soon becomes to too high to manually correct. Thus, improving speaker recognition performance reduces the required effort and costs for annotation and increases the quality of the automatically generated metadata and consequently of the search results.

Speaker recognition has been an active field of research for decades, however, a lot of mainstream speaker recognition research focuses on telephone data with its specific attributes and challenges. For telephone data, the main problem is the transmission channel which can heavily distort the speaker's voice characteristics. The main applications seem to be telephone-based access control, surveillance, intelligence, and forensics.

In order to make best use of a speaker recognition system for annotating broadcast material, it would be beneficial to adapt it to the domain's specific challenges and attributes. The main challenges are the heterogeneity of the data, the large amount of possible speakers, the acoustic environment of the recording, background music, and often the shortness of individual speaker utterances. The application, of course, is spoken document retrieval. Heterogeneity stems from the fact that broadcast material encompasses a large variety of programmes (from soap operas to news reports), but also a single news show often contains studio recorded announcements, on-site reports, soundbites, and archive material. It is harder to construct a system that handles all different types well than to devise one that specialises in only one of them. With the variety of material comes a large set of possible speakers from the different programmes, so it is easier to confuse potential speakers. Noisy acoustic environments and background music distort the speakers' voice characteristics and hamper recognition. Finally, in news shows, material from the speakers relevant to search engine users (such as celebrities and politicians) tends to consist of short soundbites embedded in longer reports, making recognition purely by voice harder because of the brevity of the test data.

However, broadcast data also offers attributes that can be exploited to benefit speaker recognition. Most importantly, speakers frequently appear in the same context. The context often contains information that helps humans identify the speakers or that may be used to narrow down the set of plausible speakers. For example, an anchorperson may announce the following speakers and give their names or mention their roles, and this information can also be exploited in automatic speaker recognition systems (see Chapter 2). Even if the names are not mentioned, the topic of a report can be used to predict which speakers are likely to appear in it — those who are considered experts in the field in question. A tennis player is much more likely to be selected to appear in a report about sports than a politician. Also, statements from a tennis player are more likely be about sports than about politics. If the assumption holds that, vice versa, a statement about sports indicates the tennis player as the speaker with a greater probability than the politician, this can be used as a speaker cue in automatic recognition systems. The

speakers' probability to appear in a recording could be gauged by their association with the topic, thus reducing the set of potential speakers. Also, in addition to a short soundbite itself, the surrounding material can be used to identify the topic with more accuracy, thus the context can be used to infer information about a short speaker utterance. Using this additional source of information together with the speakers' voice characteristics, it may be possible to increase the speaker recognition rate. There is, however, little research on topic-based speaker recognition so far. Thus, the goal of this thesis is to evaluate the usefulness of topic information for speaker recognition.

1.1 Research Questions

With the proposition of using topic information for speaker recognition, a number of questions arise that must be addressed in order to evaluate the usefulness of the approach:

Q1 How can spoken utterances' topic information be automatically captured with little manual effort, and how can it be used to learn speakers' topic preferences?

As topic information is mostly contained in the words used, automatic speech recognition can be applied to obtain word transcripts of the utterances, removing the requirement for manual transcription. Topic information must then be inferred from the words with an approach that can tolerate the transcription errors introduced by the automatic speech recognition process. The speakers' topic preferences can be modelled from speaker training utterances from the domain. The models obtained from this process can later be used to recognise a new utterance's speaker by comparing the topic information in the utterance and the preferences of all possible speakers. A score has to be derived for each possible speaker which indicates how well the topic matches the speaker. The best modelling and scoring method given the amount of training data available has to be identified. To make the use of topic-based speaker recognition practical, the topic information extraction process cannot depend on manually labelled topic training sets because those are often costly to produce or hard to obtain. Also, it would be desirable to extend the available audio training material with written text, because for most domains text data is much easier to obtain (for example on the Internet) in the large quantities necessary for training than audio material. A similar approach is common in language model training for automatic speech recognition, where written text is used instead of transcribed speech. To facilitate this, the topic information extraction process must be able to produce usable results despite the difference between spoken and written language and the transcription errors introduced by the automatic speech recognition system.

Q2 How does speaker identification based on topic cues perform and does it combine well with other cues, such as voice-based speaker recognition?

After devising a system that can identify speakers by topic information, its performance has to be

evaluated and compared to existing methods. A suitable benchmark corpus has to be selected for the evaluation. The influence of an utterances length on the performance should be gauged. Also, usually more than one speaker talks about a given topic, so it cannot be expected that the topic alone will perfectly identify the speaker. However, the topic-based speaker score may be fused with other identification systems, such as a traditional voice-based system. The question is whether the topic really adds new speaker information not captured by other systems and thus improves the overall performance.

Q3 How stable are speakers' topic preferences over time?
As the speakers' roles, jobs, and responsibilities may change, speaker topic preferences probably are not stable over time. The politicians appearing on the news can be the same for a long time but they may change departments, especially after elections. How does this affect topic-based speaker recognition? How often do speaker preferences have to be updated?

This thesis examines these questions and tries to answer them as is outlined in the next section.

1.2 Outline

This thesis is based on and uses experiences, methods, and results published in [Baum12, Baum10b, Baum10a, Baum09b, Baum09a] and is structured as follows:

Chapter 2 gives an overview on the field of speaker recognition and its mainstream research. The cues for speaker identification described in the literature are summed up: features capturing voice, prosody, pronunciation, and idiolect, as well as efforts to spot speaker names in transcriptions and related work from authorship attribution for written texts. The modelling and classification techniques employed in current speaker recognition systems are delineated: Gaussian Mixture Models — which are used in the de facto standard reference system for speaker recognition by Reynolds — as well as Support Vector Machines and Joint Factor Analysis. The methods used for feature and score normalisation and for fusion of different speaker recognition systems are briefly described.

Chapter 3 explains the two proposed methods for topic-based speaker recognition, published in [Baum12] and based on results and experiences from [Baum10b, Baum09a]. The first approach uses topic information indirectly by learning the speaker's preferred content words from example utterances, either with Log-likelihood Ratio Scoring or with Support Vector Machines based on word frequency features. Content words are extracted by removing stop words and using only the word stems, so as to retain only the linguistic units which carry topic information. This method requires no explicit topic models, but it mixes topic information with the speaker's idiolect and other speaker marker words, such as names and titles.

In order to disentangle this information, the second approach uses explicit topic models, mapping

words to topics and topics to speakers. To derive topic models for the application domain without the need for labelled training data, Latent Dirichlet Allocation (LDA), a well-known unsupervised probabilistic topic modelling algorithm, is employed. LDA learns a generative model for a domain's documents (or utterances) from an unlabelled text collection, where the topics are latent variables in the model. It was proposed for exploration of text collections because it finds the prevalent topics and the words that correspond to them. Yet, it has the advantage that topic assignments can be inferred for new documents not in the training collection. The identification of domain topics is done on written text to avoid the cost of acquiring and transcribing large amounts of audio material. Afterwards, topic assignments of the speakers' training documents can be generated and used to infer the speakers' topic preferences. These can be compared with the topics assigned to test utterances to produce a score which describes how likely the speaker is for this utterance. Three modelling and scoring methods are proposed: (1) a lazy learner using the topic assignments of all speaker training documents, (2) averaging the speakers' training material into a prototype vector, and (3) Support Vector Machines.

Chapter 4 describes the preparation of the Bundestag video corpus used to test topic-based speaker recognition. The corpus is based on experiences gained with the DiSCo corpus [Baum10a, Baum09b] and was briefly described in [Baum12]. The design of a new corpus is necessary because the standard speaker recognition benchmark corpora cannot be used to test topic-based methods, as the speakers in them were not free to choose their topics of conversation; also, these corpora do not contain broadcast material. The new corpus is based on video recordings of political speeches given in the German parliament (Bundestag). As these videos are recorded as broadcast material, and snippets from such videos are frequently used in news programmes, it should be possible to evaluate the usefulness of the proposed approaches for broadcast data on them. The corpus contains 324 hours of video recordings of speeches by 235 target speakers, covering the time from July 2009 to November 2010.

Chapter 5 details the experiments conducted and results obtained for topic-based speaker recognition and its combination with traditional voice-based and idiolectal speaker recognition. The results have in part been published in [Baum12]. The tools used for the implementation of the proposed and the baseline systems are listed and briefly described. Experiments were conducted to determine the optimal parameter settings, for example the influence of pre-processing, the best modelling schemes for both systems, the optimal number of topics, and the distance measure to use. Then, the performance of the topic-based systems was evaluated and it was found that automatically derived topic information can indeed be used for speaker identification. The influence of the length of the test utterance was examined, yielding the rule of thumb that 100 content words are needed to get good identification results. In order to see whether topic information indeed helps if added to traditional methods, the topic-based systems were fused with an idiolectal and a voice-based baseline system, and it was found that the performance

could be increased from an equal error rate of 8.6% absolute for the voice-only system to 5.3% absolute for the combination of all three. Finally, the stability of the speakers' topic preferences was estimated on a separate text corpus from the same domain which stretches over 10 years.

Finally, Chapter 6 draws conclusions and describes directions for future work.

Chapter 2

Related Work

The purpose of this chapter is to provide a brief introduction into the field of speaker recognition. As the focus of this thesis is speaker recognition in broadcast material, this introduction is concerned only with so-called text-independent speaker recognition, which does not require that the speaker reads a pre-specified prompt. This does not, however, exclude the use of automatic speech recognition software to try to recognise the words being spoken by the speaker.

All speaker recognition systems detailed in this chapter work according to the general classification scheme well-known in machine learning, illustrated in Figure 2.1: The speaker's characteristics are represented by a model or classifier which is learnt from training material. During recognition, the test material is presented to the system. It either comes from the *target speaker* or from an unknown *non-target* speaker, also called *impostor*. The test material is compared with the speaker's model and a score (for example a probability for the speaker's presence in the test material) or a decision (for example a rejection of the hypothesis that the modelled speaker is present in the test material) can be derived. If

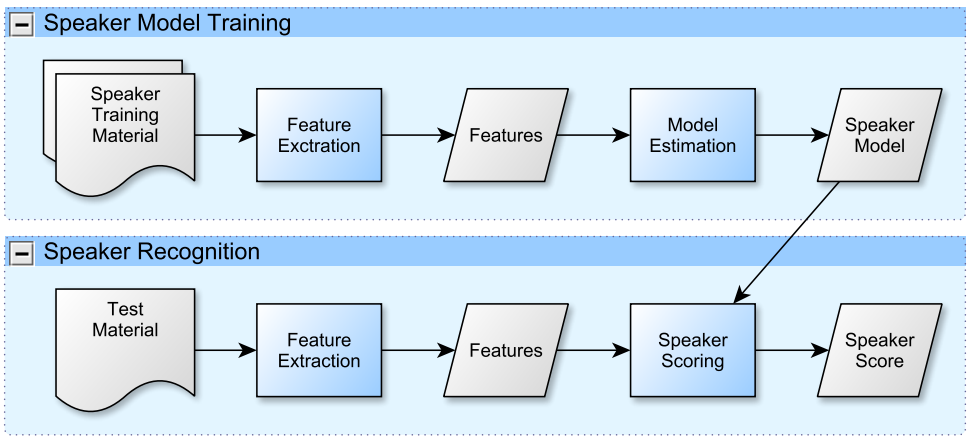


Figure 2.1: General speaker recognition system setup.

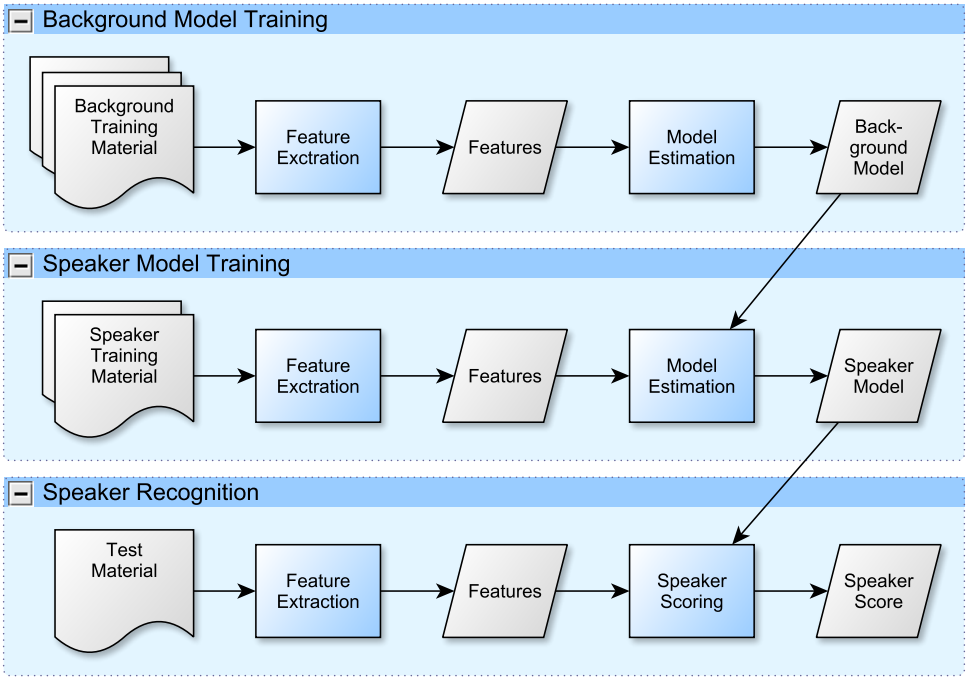


Figure 2.2: Speaker recognition system setup with background modelling.

there are multiple possible speakers from which the right one has to be identified, the necessary multi-speaker classifier is usually composed of one binary classifier for each speaker. That is, the possible speakers' models are usually stacked and the test material is compared with each of them to find the best-scoring one. The performance on this scoring / decision task can be estimated with the measures described in Section 2.1.1.

In both phases, features are extracted, which ideally would capture the relevant information in compact form and remove all unnecessary information and noise. (In practise, features are not perfect (yet) and still contain some unwanted noise.) The different cues in speech currently used for speaker recognition and the corresponding features are described in Section 2.2.

In many speaker recognition systems, however, the modelling and scoring is a little more sophisticated and includes a way of taking into account the characteristics of the "average speaker" (for the desired population of speakers), in order to find out how idiosyncratic the characteristics observed for a specific speaker are. This requires a set of sample data from the desired population, which is used to learn what is "normal". This so-called background training set ideally is very large and contains a balanced amount of material from a large number of speakers. In some systems, like the de facto standard baseline speaker recognition system (described in Section 2.3.2), the background is explicitly modelled. The speaker models are adapted from the background model and the speaker score is compared to the background score to arrive at a decision, as illustrated in Figure 2.2. The background score is important for two major reasons: First, it allows for the rejection of the speaker hypothesis when the background

score is larger than the speaker score (as the average speaker is then more likely than the hypothesised speaker). Second, it is also used for normalisation of the speaker score, as it gives a measure for how likely the test material is. The speaker models are adapted from the background model in order to have a correspondence between them and in order to reduce the required speaker training material. During scoring the background model score can then be used to normalise the speaker score, making the system more robust against variation in background noise. The most important approaches for modelling and classification are described in more detail in Section 2.3. Section 2.4 explains how features and scores for speaker recognition can be normalised, and how scores from different systems can be fused to improve performance. The next section describes how the performance of speaker recognition methods can be measured and compared.

2.1 Performance Estimation

In order to estimate the performance of a speaker recognition method, objective measures are required. A summary of the most frequently used measures is given in the next section. However, to meaningfully compare performance figures from different systems, all methods need to be tested on the same benchmark. Thus, international evaluation competitions in speech and speaker recognition have been established, using fixed benchmarks to test all competing algorithms (see Section 2.1.2 and 2.1.3). Even outside the competitions, these benchmarks have a strong influence on research in the respective fields as new methods are often measured against them.

2.1.1 Performance Measures

Well-known and established quality measures for speaker recognition are equal error rates and detection error trade-off curves, which will also be used in this thesis:

The **equal error rate (EER)** is a measure of accuracy for biometric identification systems. It is probably the most frequently used quality measure for evaluating speaker recognition, summing up the system's performance in just one number. It is usually defined for the case that a claim of identity is made by a user and this claim has to be verified by the system (binary decision). Two types of error can occur in this situation: false reject (also: miss), where a true speaker is erroneously rejected by the system, or a false accept (also: false alarm), where an impostor is mis-recognised as the target speaker. In an evaluation with n_{target} target trials and $n_{impostor}$ impostor trials, the false reject and false accept rates are calculated as $P_{fr} = \frac{n_{fr}}{n_{target}}$ and $P_{fa} = \frac{n_{fa}}{n_{impostor}}$, where n_{fr} and n_{fa} are the number of trials with false rejections and false acceptances, respectively. The equal error rate is the error rate at the system's operating point where both rates are equal, $P_{fa} = P_{miss} = EER$ [Doddington00].

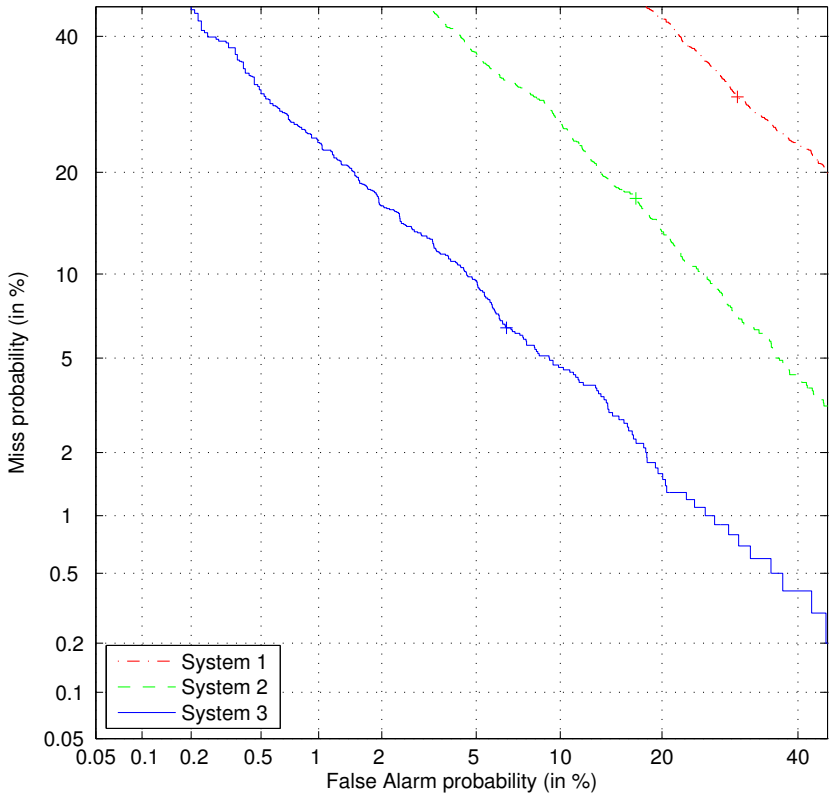


Figure 2.3: Example of a DET-plot comparing 3 systems with different performance.

For a multitarget classifier, which has to identify the correct speaker from a set of possible candidates, there is also the confusion error, when one speaker from the set produces the highest score for test material that was really spoken by another speaker from the set. Unfortunately, the probability of confusion errors strongly depends on the speaker set size [Singer04, Zigel06], which makes equal error rates for multitarget recognition hard to directly compare if the number of speaker candidates in the evaluations does not agree. However, Singer reports that the performance of a multitarget system can be predicted from the performance of the binary classifiers it is composed of, and that methods to increase the individual classifiers' performance can be directly mapped to improve multitarget detection [Singer04]. Thus, for multitarget speaker recognition it also makes sense to report the EER for the prototypical binary speaker classifier that is used as the building block. Note, that real-life systems often do not run at the equal error operating point, so the expressiveness of the value is limited for them.

In order to display system performances at different operating points, **Detection Error Trade-off curves (DET-curves)** are usually used in the literature. They are a variant of ROC (receiver operating characteristic) curves introduced in [Martin97], plotting the false acceptance (false alarm) and false rejection (miss) rates on the y- and x-axis, respectively. A normal deviate scale is used to achieve linear plots if the score values are normally distributed. This has the advantage of spreading out well-

performing systems, making them easy to distinguish. Figure 2.3 shows examples of DET-curves with corresponding equal error rates of 30% (System 1), 17% (System 2), and 7% (System 3), respectively. Note that the equal error points lie on the diagonal ($x = y$) and that a systems is better the closer its line is to the bottom left corner.

2.1.2 NIST Speaker Recognition Evaluation Benchmarks

A main driving force behind the development of new speaker recognition methods are the regular Speaker Recognition Evaluations of the National Institute of Standards and Technology (NIST). "The goal of the NIST Speaker Recognition Evaluation (SRE) series is to contribute to the direction of research efforts and the calibration of technical capabilities of text independent speaker recognition. The overarching objective of the evaluations has always been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches." [NIST09a]

The evaluations have run annually from 1997 to 2006 and have switched to biannually since. In 2010, more than 50 institutions and groups of institutions participated, comparing their systems on the task "to determine whether a specified speaker is speaking during a given segment of conversational speech" [NIST10a, Section 2.1]. As the specified speaker is described by one or several recordings of his or her speech, the task can also be expressed as "Is the speaker in the training material the same as in the test material?" There were nine separate task conditions, differing in the length of the available speaker training and test material. Only the "core" condition was compulsory for all participants while the others were optional. In the core condition, each training and test recording was either a five minute excerpt from a telephone conversation (with the two speakers given in two different channels) or a 3 to 15 minute excerpt from an interview held in a small room and recorded with microphones. The other conditions included training and test recording lengths ranging from 10 second segments of telephone conversations to sets of 8 telephone conversations involving the speaker in question. In 2001, an extended data task was introduced which offers more training and test data per speaker in order to stimulate the development of idiolectal characteristics for speaker detection [NIST01]. In this task, all conversation sides of the target speakers available in the Switchboard corpus (up to 16) could be used for training, amounting to up to 1 hour of training material per speaker. The tests were done on whole conversation sides, which average to 2.5 to 3 minutes of audio [Switchboard]. This extended task has sparked research into "high-level features" (see Section 2.2).

The different datasets composed for each of the NIST evaluations are often used as benchmark corpora by researchers to compare new approaches to established speaker recognition methods. The datasets are made up of data from several corpora, in particular Switchboard [Switchboard], Fisher

[Cieri04b], and Mixer [Cieri04a]. The benchmarks for the evaluations from 2000 to 2005 are available through the Linguistic Data Consortium (LDC)[1], the newer benchmarks have not yet been publicly released, only to the institutions that participated in the corresponding evaluations have access to them. The data consists of telephone conversations and interviews recorded in small rooms. Depending on the recognition task, the training and test material is between 10 seconds and 15 minutes long. In the 2010 benchmark, there are 442 speakers, 15,834 target trials and 594,915 nontarget trials for the core condition. Each conversation was assigned a topic (the same topic for all conversations recorded on that day) picked from a prepared topic list, and speakers were asked to talk about this subject ([Cieri04b, Cieri04a, Doddington01]).

The importance of the NIST benchmark data seems to drive the research field towards approaches that can be evaluated on NIST data and tends to marginalise approaches that can not. The application scenario for speaker recognition methods tweaked to this data most likely is language independent telephone surveillance and forensic speaker recognition, not spoken document retrieval. Thus, the research efforts related to the evaluations is more focused on speaker characteristics that are language independent and hard to consciously control and change, especially voice-based features. This may explain why high-level features for speaker recognition have not received as much attention as low-level features, and also why topic-based speaker recognition is yet unexplored.

2.1.3 Automatic Speech Recognition Benchmarks

The National Institute of Standards and Technology (NIST) also repeatedly organised international speech recognition evaluations, starting in 1985 [NIST09b] and resulting in the creation of the Rich Transcription evaluation series in 2002 [NIST10b]. In this series, speech-to-text transcription and related systems (for example voice activity detection and speaker diarization) have been evaluated on data from various domains, including telephone, broadcast news, interview, and meeting data.

In 1995, NIST started evaluating speech recognition on broadcast material and created the HUB 4 broadcast news benchmark evaluations, which ran annually until 1999. These benchmarks cover data from radio and television broadcasts, over 200 hours in total, mainly in American English (but also including Spanish and Mandarin). Over the years, the word error rates (WERs) of the best performing systems in the evaluations decreased from 27.0% for 1995 to 13.5% for 1998. However, this variation also partly stems from the different proportions of "easy" studio-recorded material in the different data sets [Pallett02]. The HUB 4 benchmarks are available from the Linguistic Data Consortium [2] and are frequently used to measure English ASR systems aimed at the broadcast domain.

[1] `http://www.ldc.upenn.edu/` last accessed: 2013-02-25
[2] `http://www.ldc.upenn.edu`, last accessed: 2013-02-25

For French, a broadcast news transcription evaluation benchmark was created in the ESTER project funded by the French Ministry of Research and New Technologies. This ESTER benchmark contains 100 hours of manually annotated and 1,677 hours of non-transcribed radio and television data from French language broadcasting stations [Galliano06]. It is available from the European Language Resources Association [3] and is often used in evaluating French broadcast ASR system.

2.2 Cues and Features for Speaker Recognition

Most systems which are currently successful in NIST evaluations fuse the output of several speaker recognition methods to improve performance. At their core, many rely on voice-based (spectral / cepstral) approaches, but since the extended data task in the 2001 evaluations sparked more interest in "high-level" features, other cues for speaker recognition have successfully been tested. They contribute valuable information in the fusions and improve the overall system performance significantly.

Ideal speaker recognition features should:

- be very similar for material from the same speaker and very different for material from different speakers
- occur frequently in natural speech
- be robust against noise and channel effects
- be robust against manipulation by the speaker and against impersonation by an impostor
- be easy to extract
- not require much audio material.

No current feature extraction method yields such perfect features (yet) and it is debatable whether it is even possible to create one single such method. However, humans seem to use more than one set of cues to correctly identify a familiar speaker, thus mitigating imperfect information — and so can automatic systems [Reynolds03b]. Figure 2.4 shows the different levels of speaker information present in speech, which can be fused to improve a speaker recognition system's performance considerably. The corresponding features are described in the following sections. For an orientation on how well the features perform individually, equal error rates (EERs) for representative systems are given at the end of each section. They are also summed up in Table 2.1 for convenience.

It appears that the further up in the pyramid the cue is, the more sparse is the research on features capturing these cues. This is understandable as the amount of required training data becomes larger in

[3] `http://catalog.elra.info/product_info.php?products_id=999`, last accessed: 2013-02-25

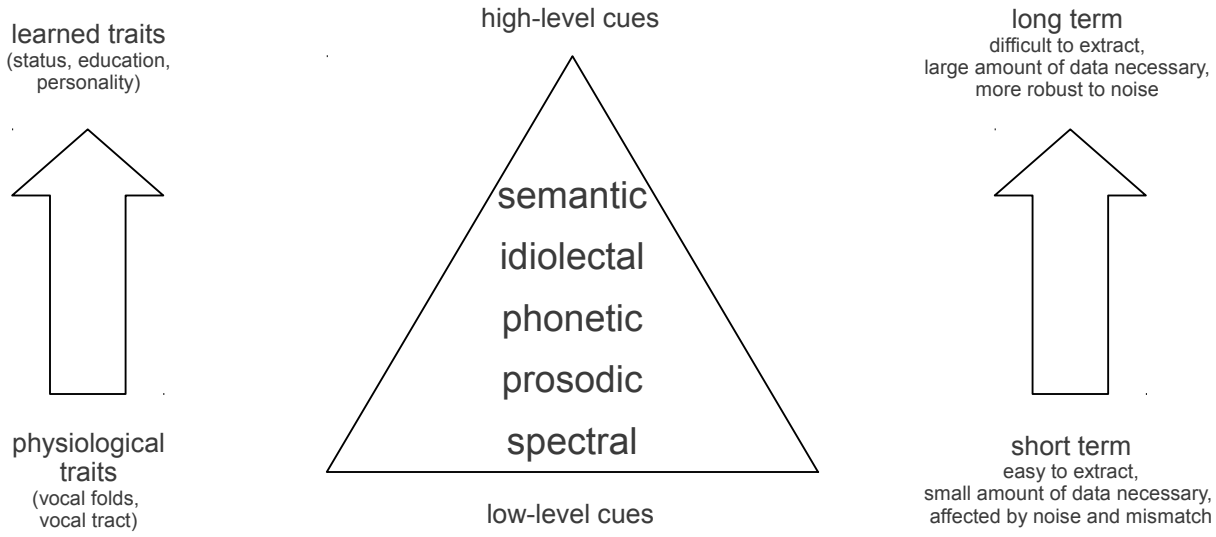


Figure 2.4: Levels in speech giving different speaker recognition cues (adapted from [Reynolds03c]).

feature type	test corpus	reported EER	reference	described in
spectral	2006 NIST SRE	4.6%	[Kenny08]	Section 2.2.1
prosodic	2006 NIST SRE	11.0%	[Kockmann10]	Section 2.2.2
phonetic	2003 NIST SRE	8.5%	[Hatch05]	Section 2.2.3
idiolectal	Fisher subset	27.4%	[Shriberg05]	Section 2.2.4

Table 2.1: Comparison of the different feature types. Training and test material is 2.5 min respectively in all cases.

this direction, and lack of training data is a main restriction for speaker recognition systems in practise. This may also be the reason why "semantic" speaker recognition has not been investigated deeply (to the author's knowledge).

The rest of this section is dedicated to giving an overview of the existing features, starting at the bottom of the pyramid, moving up.

2.2.1 Voice-based / Spectral Speaker Recognition

The most low-level — and universal — cue for speaker recognition is the speaker's voice, which can be captured by modelling spectral / cepstral features[4]. This has several advantages: Voice recognition is mostly language independent, especially compared to higher level cues. The features are based on very short segments of speech (typically in the order of 10-30 milliseconds), thus requiring less training and test audio material than the higher level features with longer segments. Also, the speaker's voice is determined in large part by physical traits (anatomy of the vocal tract) and it is hard to consciously alter these characteristics, which makes the method robust against manipulation. Thus, most successful

[4]In this thesis, voice-based features will be often referred to as "spectral features", as is common in speaker recognition literature when distinguishing between high- and low-level features, even if this is somewhat imprecise.

speaker recognition systems include (at least) a voice-based subsystem.

The most commonly used voice-based features for speaker recognition mainly stem from speech recognition — which may be surprising because speech recognition features ideally should not extract speaker specific characteristics from the audio signal. Common speech recognition features, however, still contain speaker information and are successfully used for speaker recognition. Although attempts have been made at creating specific speaker recognition features (e.g. [Miyajima01]), the traditional speech features still prevail in the successful speaker recognition systems. Among the most commonly used features are Mel-frequency cepstral coefficients (MFCCs) [Davis80], linear predictive cepstral coefficients (LPCCs), and perceptual linear prediction (PLP) coefficients [Hermansky90], which all are different ways of describing the spectral content of an audio signal.

Apparently, the selection of appropriate channel compensation methods is more important than which of those features to use [Kinnunen10]. This also points to the main drawback of spectral features: they are vulnerable to channel variations and noise [Reynolds95b, Reynolds03a] because these influences distort the spectral characteristics of speakers. Thus, in the context of speaker recognition on broadcast data, the heterogeneity the material is problematic for them. A number of ways for dealing with this weakness have been developed, ranging from special modelling methods (see Section 2.3) to normalisation techniques (see Section 2.4). Fusion (also see Section 2.4) of spectral with higher-level approaches is another way that is used by many systems successful in the NIST evaluations today.

The *de facto* reference method for (spectral) speaker recognition is the so called GMM-UBM approach by Reynolds, described in [Reynolds00], which employs MFCC features. The modelling used in this system will be explained in more detail in Section 2.3.

To give an estimate of how well spectral systems perform, here are some examples of error rates: [Campbell06b] reports an equal error rate of 5.68% for the standard GMM-UBM system on the 2005 NIST SRE corpus, with approximately 20 minutes of training material (from 8 telephone conversations) per speaker and 2.5 minutes of test material per trial. With more elaborate modelling and score normalisation (see Sections 2.3 and 2.4), better performance can be achieved: in [Kenny08] an EER of 4.6% for the female speaker subset of the NIST SRE 2006 data (2.5 minutes of training and test material, respectively) is reported. For the test setup with 20 minutes (8 conversation sides) of training material and 2.5 minutes of test material, they report an EER of 2.3%.

2.2.2 Prosodic Speaker Recognition

Prosody, in linguistics, is the intonation, rhythm, and stress of an utterance, stemming from variation in pitch, segment length, loudness, and formant frequencies. Pitch features were tested early on for speaker

recognition [Atal69], and later on explored in more detail, for example in [Carey96, Sönmez98, Weber02, Peskin03]. They were successfully fused with spectral approaches in order to make the resulting systems more robust. More recent approaches can be found for example in [Shriberg05, Dehak07, Kockmann08, Kockmann10, Ferrer10b].

Basic pitch features encompass statistics about the fundamental frequency (pitch), such as maximum, minimum, mean, variance and the derivatives and slope. The fundamental frequency itself can be extracted with auto-correlation algorithms such as [Boersma93] and [Cheveigné02].

When using alignment information from a word or phoneme recogniser, statistics like phone, syllable, word, and pause durations can be used to capture speaking style.

A more sophisticated approach, trying to capture all kinds of prosodic cues, is described in [Shriberg05]. It is a syllable-based approach, where numerous duration, pitch, and energy statistics of syllables are collected. These are discretised by binning, combined into n-grams, and the occurrence counts of the n-grams are fed into a Support Vector Machine for speaker classification.

More complex approaches to modelling the temporal trajectory of the pitch are compared in [Kockmann10]: Pitch values are extracted, utterances are segmented into syllables, and the pitch contour over the syllables is modelled. For pitch extraction, algorithms based on auto-correlation of energy seem to perform best. Syllable segmentation is best performed by a large vocabulary continuous speech recogniser (LVCSR). The temporal trajectory of the pitch can be modelled by a Discrete Cosine Transform or a Legendre polynomial. Taking the coefficients of the DCT or polynomial, the pitch contour of each syllable can be rendered as a prosodic feature vector. These vectors can then be used for speaker modelling, for example with the methods described in Section 2.3.

For an impression of the equal error rates achieved: In [Shriberg05], an equal error rate of 11.18% for the best performing prosodic system on a self-composed test set made from the Fisher corpus [Cieri04b] is reported, with 2.5 minutes of training and test material per speaker. In [Kockmann10], the best system has an EER of 11% on the NIST SRE 2006 data with 2.5 minutes of training and test material per speaker, respectively.

2.2.3 Pronunciation-based / Phonetic Speaker Recognition

A speaker's idiosyncratic pronunciation may include, for example, omitting certain phones at the end of some words, using a specific set of phones due to their dialect or accent, or co-articulating some phones in a particular manner. What ever the particulars of the speakers' deviation from standard pronunciation, it can be used to identify them. As this approach relies on phone symbols instead of spectral vectors, it is believed to be more robust against channel mismatch than spectral speaker identification.

Pronunciation-based speaker recognition requires a phone recogniser to produce a transcription of the sounds uttered by the speaker. A phone recogniser is an important part of an automatic speech recognition systems, where it is combined with a dictionary and a grammar or language model to incorporate information about the uttered sounds with knowledge about word pronunciation and frequently used phrases. However, the systems described in this section employ phone recognisers that do not use a dictionary, language model or a grammar, which means that they just rely on the acoustic, spectral information in the audio signal to identify phones. They do not use this additional information to avoid language dependence. Most of the systems use the Zissman phone recogniser [Zissman96], which is based on spectral features (MFCCs) as input features and models each phone as a sequential three-state Hidden Markov Model (HMM). Note that although the spectral features are used for the phone recognition process, they are *not* directly used for speaker recognition.

The first approach to automatic phonetic speaker recognition is described in [Andrews01, Andrews02]: Six language dependent phone recognisers, each trained with audio material from one of six languages[5], are used to create six language dependent speaker identification systems. In each of the systems, the phone transcription is first converted into sequences of n-grams (in this case n=3). Then the frequencies of each n-gram in the speaker training material and the background training material are compared. The ratios of speaker and background n-gram frequencies capture whether a speaker uses a phone n-gram more or less often than normal. During recognition, these ratios are summed up for the phone n-grams present in the test material. This yields a likelihood ratio indicating how much more likely the test material comes from the speaker than from another person. In the end, the scores of the six language dependent subsystems are fused together by linear combination, thus producing a score for a language independent system.

In [Jin03] the same six language-dependent phone recognisers are used for a different approach: instead of relying on n-grams in the time dimension (capturing the chronological co-occurrence of phones), the system creates n-grams across the language specific phone transcriptions. This means that the n-grams capture which phones are output by the different phone recognisers at the same point in time. The modelling is done as in the previous method, yielding likelihood ratios of speakers against the background. When fused with the previous approach, this significantly reduces the equal error rate.

An effort to model the differences between the standard and the actual pronunciation of words is made in [Klusacek03]. The approach relies on both a phone and a phoneme transcription produced by an open-loop phone recogniser and an automatic speech recognition system with a language model and a pronunciation dictionary, respectively. With both transcriptions as a basis, conditional probabilities of the phone given the phoneme (the realisation given the intention) are produced for the speakers

[5] English, German, Hindi, Japanese, Mandarin, and Spanish

and the background. The ratio of both probabilities gives an indication of how specific this pair is for the speaker. During recognition, these ratios are summed up for the phone/phoneme pairs in the test material, yielding a likelihood ratio for the speaker compared to the background. Keep in mind that this system is language dependent because it relies on language dependent ASR to produce the phoneme transcription.

A better way of modelling the phone n-gram frequencies is explored in [Campbell03b], where Support Vector Machines (SVMs) are used instead of simple likelihood ratio tests. Relative phone n-gram frequencies for each audio file are extracted, as in the first approach ([Andrews01, Andrews02]), and treated like term frequencies in a text classification scheme. They are weighted with a log-likelihood ratio term weighting, which captures how often a term appears in the background material ($weight = 1/\sqrt{p(term|background)}$). The term frequency vectors for the speakers' training documents are then used to train the speaker SVM models. The term frequency vectors of the test documents are fed into the SVMs for classification during the recognition phase. Just as before, 5 different speaker recognition subsystems are built with the phone streams from 5 different language-dependent phone recognisers. The 5 subsystem scores are fused by linear combination into a language-independent system score, which outperforms the individual subsystems. More importantly, the SVM approach significantly outperforms the likelihood ratio tests.

The SVM modelling is also used in [Hatch05], but here the n-gram frequencies are computed differently. All previous approaches have used 1-best phone transcriptions: the phone recogniser's best hypothesis on which phones are present in the audio signal. An alternative is to use phone lattices, containing also other hypothesis and their probabilities in a lattice graph produced by the phone recogniser. Information about the phone n-gram frequencies is extracted from the phone lattices: the count of the n-grams in each of the phone sequences in the lattice is multiplied by the posterior probability of the sequence. The sum of those weighted counts is then called the expected phone n-gram count, giving a probabilistic measure of the number of occurrences of the n-gram. The expected counts are converted to relative frequencies and then modelled with an SVM like in [Campbell03b].

For an estimate of the performance of phonetic speaker recognition, take the lattice-based SVM system from [Hatch05]. Tested on the NIST SRE 2003 Extended Data with 20 minutes (8 conversation sides) of training material and 2.5 minutes (1 conversation side) of test material per speaker, it has an EER of 2.0%. For 2.5 minutes of training and test material it still achieves an EER of 8.5%.

2.2.4 Idiolectal / Word-based Speaker Recognition

Idiolect, for the purpose of speaker recognition, is the idiosyncratic word usage of a speaker: frequently used words, idioms, and mannerisms. As the units of speech in this approach are words, it requires an automatic speech recognition system for the language in question. While this may pose a problem for language independent speaker recognition, like in the NIST SRE, for a spoken document retrieval system where ASR is applied to all audio documents anyway, word-based features can bring valuable information to a speaker recognition system.

Idiolectal speaker recognition was first explored in [Doddington01] on manual transcriptions of the Switchboard corpus [Switchboard]. To extract idiolectal features, the words in an utterance are combined into n-grams, and the n-gram frequencies are computed to capture information about frequently used and speaker-idiosyncratic phrases. For training a speaker model, the log likelihood ratio of the speaker likelihood and the background set likelihood of each n-gram type is computed (the likelihoods are approximated by the relative n-gram frequencies in the speaker training and background material). This yields a measure of how idiosyncratic the n-gram is for the speaker. For testing against a speaker model, all log likelihood ratios for the n-grams in the test data are summed up (weighted according to the number of occurrences of the n-gram in the test material) and normalised with the length of the test. This yields an average test n-gram log likelihood ratio of the speaker against the background. The best results are achieved using only the 2,000 most frequent word bigrams in the Switchboard corpus, thus focusing on commonly used word patterns to identify speakers. Note that this approach is aimed at stylistic phrases (for example "how shall I say", "you bet", or "in terms of"), not at words specific to favourite topics. This is because in the Switchboard corpus used for the study, speakers where required to talk about preset topics and were not free to choose according to their own preferences.

In [Shriberg05], a word n-gram SVM system is briefly described, which uses Support Vector Machines (SVMs) with linear kernels to discriminatively model the relative word n-gram frequencies. The words themselves are the output of SRI's English conversational telephone speech recognition system, not manual transcriptions, as in the previous study. All n-grams up to $N = 3$ which occur more than once in the background set are used, yielding roughly 150,000 different n-grams.

This resembles text classification with SVMs [Joachims98] if the speakers are used as the categories, but differs in a number of aspects: error-prone ASR transcripts are used instead of written text, n-grams instead of single words; no stemming and stop-word-removal is applied, and a different term weighting scheme for normalising the n-gram counts is employed (see Section 3.4.2 for an explanation of the different term weighting schemes). A similar approach was used by the same group in their entry into the NIST SRE 2010, using only the 9,000 most frequent bi- and trigrams from the training data [Ferrer10a].

For an impression of the achieved performance and required audio material: The experiments in [Doddington01], conducted on the NIST SRE 2001 extended data task, produce an equal error rate of approximately 10% for speaker models trained on 8 Switchboard conversation sides (approximately 20–25 minutes of speech) and tested on approximately 1 Switchboard conversation side (approximately 2.5–3 minutes of speech). If the speaker models are trained and tested on only on one conversation side, the system achieves an equal error rate of approximately 25%. Keep in mind that this system uses manual transcripts, thus simulating perfect ASR. In [Shriberg05] the word n-gram SVM system has an equal error rate of 27.4% on the self-composed test set made from the Fisher corpus [Cieri04b]. Here, the length of the training and test files is restricted to approximately 2.5 minutes per speaker each.

2.2.5 Naming unknown speakers

Another approach to use ASR transcripts to gain speaker information is to search it for possible speaker names. In a domain like broadcast news, where speakers are often introduced or introduce themselves, linguistic patterns like "I am <name>" or "joining us now is <name>" can be identified in the transcript and exploited. Thus far, this approach was applied for speaker diarization where no pre-trained speaker models exist and the task is to tell apart the different speakers in a broadcast. Spotting names in introductions makes it possible to assign these names to the unknown speakers [Canseco-Rodriguez04, Canseco05, Mauclair06]. As the task is different, there are no equal error rates in these publications. However, for the English Hub4-E and French ESTER broadcast news corpora, [Canseco05] and [Mauclair06] claim a speaker name identification accuracy of approximately 80%, where naming is most successful when the speakers introduce themselves (and not others).

2.2.6 Authorship Attribution for Written Texts

Parallel to speaker identification in spoken documents, authorship attribution for written text (using statistical methods) has been a field of research for decades, becoming especially active in the last decade, as large text corpora are now easily accessible through the Internet (Usenet, e-mail, etc.) [Stamatatos09]. As with speaker identification, authorship attribution is applied in intelligence, homeland security, forensics and other crime-related fields [Abbasi05, Chaski05]. It also has been tested on newspaper articles [Diederich03] and traditionally, literary texts have often been analysed for authorship, e.g. in [Burrows02] or the famous analysis of the Federalist papers in [Mosteller64].

In contrast to speaker recognition, no voice information is available in authorship attribution, thus authors have to be identified by features derived from words only. Most approaches identify writers with so-called stylometric features, which aim at capturing idiosyncratic writing style, independent of

the topic discussed. Common features are sentence lengths, punctuation mark frequencies, word length, word and word-n-gram frequencies, character and character-n-gram frequencies, vocabulary richness and part-of-speech frequencies [Stamatatos09]. While many of these features have their parallels in speaker identification, others are not (yet) directly applicable to spoken documents because of restrictions of the automatic speech recognition methods available. For example, analysis of sentence lengths and punctuation mark frequencies would depend on the speech recogniser's ability to correctly produce punctuation, which still is a difficult problem in ASR. However, pauses in speech, which are used for generating punctuation in ASR [Shen09], can also be directly used for speaker identification with prosodic features, such as in [Shriberg05]. A speaker's vocabulary richness (measured by the number of different words used) extracted from ASR transcripts may be limited by the size and content of the speech recogniser's and not the speaker's dictionary.

Word occurrences of very frequent words (stop words) are reported as a very effective author identification feature, for example in [Argamon05]. Correspondingly, word-n-gram occurrences of frequent words have been used for idiolectal speaker recognition (see Section 2.2.4), yet their reliability may be reduced by ASR recognition errors, which are common especially on the short function words. Frequently, stylometric features are designed and used to avoid topic-dependence, in order to be able to recognise authors even when they discuss other topics than in the training material. However, if topic preferences of authors are present for the application domain (such as in personal on-line communication), using features capturing topic seems to significantly improve the recognition accuracy for authorship attribution [Goldstein-Stewart09, Zheng06]. Thus, in domains such as broadcast news, where speaker-specific topic preferences can be assumed, it appears very sensible to make use of them. A remaining problem seems to be the automatic selection of topic features for the domain: In [Goldstein-Stewart09, Zheng06] very few (5 and 11) marker words were manually selected as indicators for topic content, resulting in very basic features. Stamatatos remarks: "It remains unclear how to select such features for a given text domain." [Stamatatos09].

For classification, Support Vector Machines are popular in authorship attribution because they work well on high-dimensional data: They can efficiently process the sparse vector space model representations of the texts, where each distinct word in the text collection is a feature dimension [Diederich03, Argamon05, Goldstein-Stewart09].

A number of evaluation corpora for authorship attribution from various domains (literature, newspaper articles, e-mails, etc.) have been built, and corpora designed for other purposes have been used to study authorship attribution. However, the various corpora were often used only in a few studies, and although efforts were made to establish a common benchmark [Stamatatos09, Stamatatos01], which has been used by several groups, it seems not to have reached the prevalence of, for example, the NIST SRE

benchmarks. To get an idea of the performance that can be achieved in authorship attribution, a few examples with a short description of the respective corpus is given here:

In [Goldstein-Stewart09], a corpus of personal communication in different modes is used. It contains text and manually transcribed speech from essays, emails, blogs, chatlogs, interviews, and discussions from 21 subjects, with at least a few hundred words per person and mode. The best person identification accuracy of 82% averaged over all modes is achieved with a combination of lexical, syntactic, structural, and content-specific features and Support Vector Machine classifiers.

In [Zheng06], the corpus consists of newsgroup messages linked to cybercrime. It contains an English and a Chinese part, with 20 authors in each part, at least 30 messages per person, and an average of 169 words and 807 words per message, respectively. The best person identification accuracies of 97.69% (English) and 88.33% (Chinese) is again achieved with a combination of lexical, syntactic, structural, and content-specific features and Support Vector Machine classifiers.

In [Diederich03], the domain is newspaper articles with at least 200 words from 7 authors, word frequencies and Support Vector Machines are employed, and the accuracy is mentioned as being 60-80%.

In [Argamon05] the corpus is composed of chapters of novels by 8 authors with at least 5 chapters per author and a few thousand words per chapter. An accuracy 99% is achieved, with frequencies of frequent words as features and Support Vector Machine classifiers.

2.3 Speaker Modelling and Classification

After feature extraction, the second crucial component in a speaker recognition system is the modelling applied to capture the speakers' characteristics, and the corresponding classification method. A small number of techniques are usually used for most of the features detailed in the previous section, among them Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) and log-likelihood ratios. Characteristically, speaker recognition systems include a way to take into account the background, the "average speaker" (see Figure 2.2). Therefore, when setting up a system, it is important to have sufficient data not only for the training of speaker models but also for the background model — usually this is a substantial amount.

2.3.1 Log-likelihood Ratios

Log-likelihood ratios are frequently used in speaker recognition to take into account the background, the "average speaker". This is done by computing the ratio of the likelihoods of the hypotheses that the audio material was produced by the hypothesised speaker (H_0) or by the background population (H_1),

respectively. For computational and numerical reasons[6], it is usually the logarithm of the likelihoods that is used. Thus, the average log-likelihood for a test segment x (with feature vectors $x_1, ..., x_T$) is

$$\Lambda(x, \lambda_{spk}, \lambda_{bg}) = \frac{1}{T} \sum_{t=1}^{T} \{\log p(x_t|\lambda_{spk}) - \log p(x_t|\lambda_{bg})\}, \tag{2.1}$$

where λ_{spk} is the speaker and λ_{bg} the background model and T is the number of feature vectors. How $p(x_t|\lambda)$ is calculated depends on the modelling approach, of course.

If Λ is close to 0, the hypothesised speaker is approximately as likely as the average person. If $\Lambda < 0$, then the background is more likely than the speaker. If $\Lambda \gg 0$, the speaker is much more likely than the average person. In order to quantify "much more", a threshold α can be defined so that the speaker hypothesis is only accepted if $\Lambda > \alpha$.

A simple example are the log likelihood tests for word or phone n-grams in Sections 2.2.3 and 2.2.4. Here, the n-grams n_i are the features and their relative frequencies in the speaker and background data, serve as the probabilities of the feature given the speaker or background, $p_{spk}(n_i)$ and $p_{bg}(n_i)$. In the easiest case, the sum of the feature log-likelihood ratios is normalised with the number of n-grams in the test material, N:

$$\Lambda(x) = \frac{1}{N} \sum_{i=1}^{N} \{\log p_{spk}(n_i) - \log p_{bg}(n_i)\}, \tag{2.2}$$

But, of course, log-likelihood ratios can be used with any sort of model that outputs probabilities for the speaker and the background — for example with Gaussian Mixture Models (see the next section).

Likelihood ratios are the preferred form of statements in forensic speaker recognition [Champod00] because they enable the forensic expert to make a statement about the relative likelihood of a speaker, without taking into account the speaker's prior probability. That is, they can state how likely the speaker is in comparison to the background population without having to consider the other evidence presented in court. This is regarded as the jury's or judge's task. For speaker recognition in broadcast data, this means that the speakers' relative likelihood can be assessed without having to know their prior in advance. If the prior is known, it can be easily multiplied with the likelihood ratio (or added to the log-likelihood ratio) to derive the odds of the speaker being present in the audio material.

2.3.2 Gaussian Mixture Models

Amongst others, the standard reference for speaker recognition systems, [Reynolds00], uses Gaussian Mixture Models (GMMs) for modelling. As this system is so pervasive in the literature, the approach

[6] Computational efficiency may be improved by using additions instead of multiplications — this substitution can be made when the operands are mapped to their logarithms. Also, using logarithmic values avoids arithmetic underflow when very small probabilities are computed.

will be summed up here. As features, Reynolds uses MFCCs, but GMMs can be used with other spectral or high-level features as well.

A Gaussian Mixture Model λ (see, e.g., [Bishop95, Section 2.6]) approximates a probability density function $p(x|\lambda)$ with a linear combination of M component Gaussian densities $p_i(x)$ $(i = 1, ..., M)$:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, p_i(x), \tag{2.3}$$

where w_i are the component weights that sum to 1 ($\sum w_i = 1$).

The component densities $p_i(x)$ in d-dimensional feature space are defined as

$$p_i(x) = \frac{1}{(2\pi)^{d/2} \, |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2}(x - \mu_i)'(\Sigma_i)^{-1}(x - \mu_i) \right\}. \tag{2.4}$$

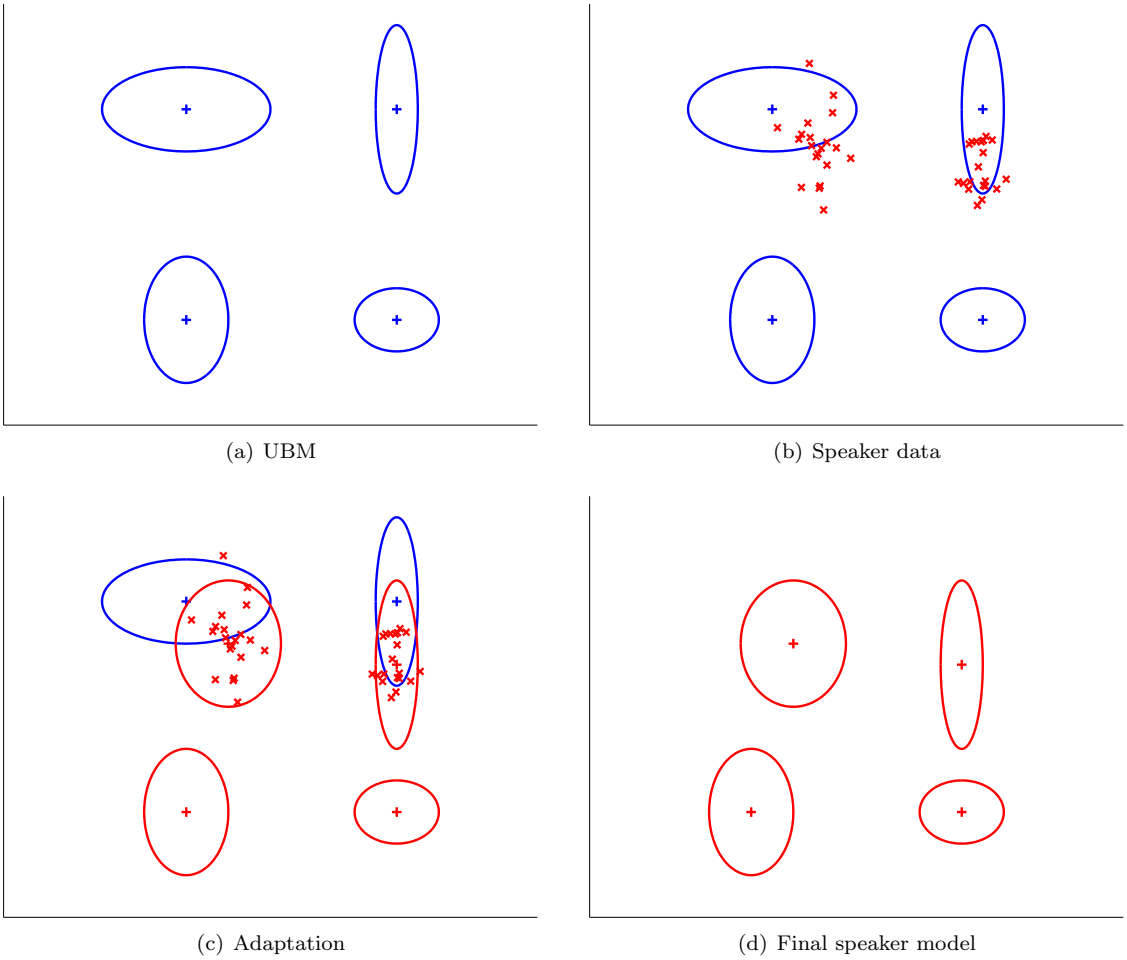
Thus, the components are described by their weight w_i, their mean value μ_i, and their covariance matrix Σ_i. These parameters can be trained with the expectation-maximisation (EM) algorithm ([Bishop95, Section 2.6]).

The goal in Reynolds's approach is to output log-likelihood ratios (see the previous section), so it needs to model not only the speaker but also the background population (or average speaker). For this, a *universal background model* (UBM) is trained with a large amount of speech data stemming from a large number of speakers. Usually tens or hundreds of hours of training material are used for this purpose.

Then, the speaker models are generated: instead of training them just from the speakers' training data, the data is used to adapt the background model into speaker models (cf. Figure 2.2). This has several advantages: It enables the system to train good speaker models with little training data. Also, it introduces a correspondence between speaker and background model that apparently increases the system's recognition performance and allows faster computation of scores during recognition (see [Reynolds00, Section 3.4] for details). To derive the speaker models, the *maximum a posteriori* (MAP) method is used: First, sufficient statistics for the weight, mean, and variance parameters are computed from the speaker's training data x. Then, these are used to update the UBM's weight, mean, and variance parameters by computing weighted combinations, with weight α. For example, the updated mean value is computed as:

$$\hat{\mu_i} = \alpha_i E_i(x) + (1 - \alpha_i)\mu_i, \tag{2.5}$$

where μ_i is the UBM mean value and $E_i(x)$ the mean statistic from the speaker data for the ith Gaussian component. The combination weight α determines the emphasis that is put on the speaker's training data (in contrast to the UBM): if $\alpha \to 1$, the influence of the training data is high, if $\alpha \to 0$, the influence of the UBM is high. In Reynolds's system, an individual, data-dependent weight α_i is used for



(a) UBM

(b) Speaker data

(c) Adaptation

(d) Final speaker model

Figure 2.5: MAP-Adaptation for a 4-component GMM in 2-dimensional space: The UBM (2.5(a)) is adapted with speaker data (2.5(b)): The parameters of the components closest to the data are adapted (the upper mixtures), the other components are taken from the UBM (the lower mixtures) (2.5(c)). A correspondence between the UBM and the resulting speaker model is still visible (2.5(d)).

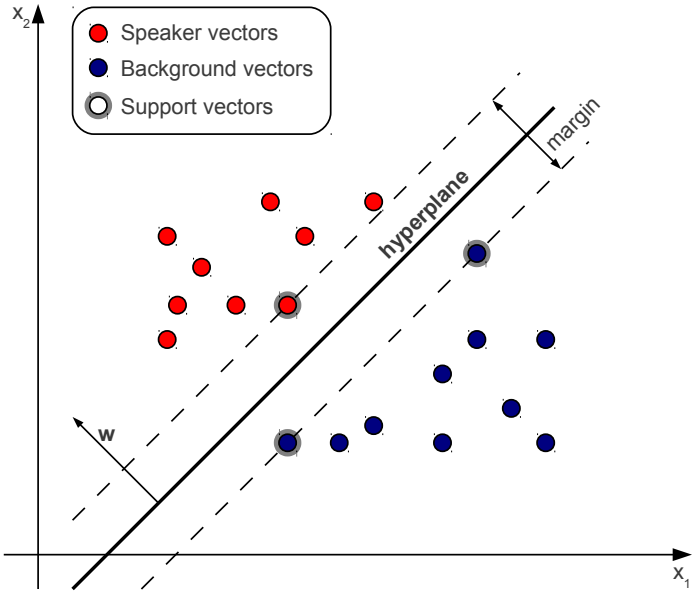


Figure 2.6: Support Vector Machine example in 2-dimensional space.

each component i: the components that fit the speaker's data better are adapted more, those who do not fit the data well are left as in the UBM. This process is illustrated in Figure 2.5.

After the speaker models are trained, they can be used for speaker recognition together with the UBM. For test features x, the speaker's score is given as the log-likelihood ratio between speaker and background model,

$$\Lambda(x) = \log p(x|\lambda_{Speaker}) - \log p(x|\lambda_{UBM}). \tag{2.6}$$

Reynolds's approach is often referred to in the literature as the "GMM-UBM" system.

2.3.3 Support Vector Machines

Apart from Gaussian Mixture Models, a second classification approach is frequently employed in speaker recognition: Support Vector Machines (SVMs). They have been used to model not only spectral [Campbell06a] but also prosodic [Shriberg05], phonetic [Campbell03b, Hatch05], and idiolectal [Shriberg05] features.

GMMs model the probability density of a speakers training features, thus describing the speaker by the features they usually generate. In contrast, SVMs [Cortes95] model the decision boundary between two classes, the speaker and the background population. The boundary is a hyperplane that separates the feature vectors from both classes from each other (see Figure 2.6). It is learnt from positive examples – the speaker's features – and negative examples – the background data features. More specifically, SVMs find the optimal separating hyperplane that maximises the margin between the vectors of the two

classes.

The distance of a feature vector $\mathbf{x}$ from the hyperplane is given by $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$, where $\mathbf{w}$ is the normal vector of the hyperplane and b determines its offset from the origin. Each training vector's class is indicated by a class label $y \in \{-1, 1\}$. The margin is given by $\frac{2}{||\mathbf{w}||}$, so in order to maximise it, $\mathbf{w}^2$ must be minimised under the constraint $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \forall i$. The constraint ensures that the vectors of both classes are separated by the hyperplane and do not fall into the margin. It can be shown that the hyperplane is determined only by the vectors closest to the boundary, the so-called *support vectors*.

For the case that the training data cannot be separated without error, a soft margin hyperplane can be used, allowing a minimal number of errors. This entails introducing slack variables ξ_i, which measure the classificaiton error of the feature vectors $\mathbf{x}_i$, into the minimisation term. The term is changed from $\mathbf{w}^2$ to $\frac{1}{2}\mathbf{w}^2 + C \sum_{i=1}^{n} \xi_i$, where C is a constant, the cost assigned to a missclassification. The constraint is changed to $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i, \forall i$.

For classes that are not linearly separable, the kernel trick can be used to project the feature vectors into a higher dimensional space, where they can be linearly separated. In general form, the distance of a feature vector $\mathbf{x}$ to the hyperplane is given by $f(\mathbf{x}) = \sum_i y_i \alpha_i K(\mathbf{x_i}, \mathbf{x}) + b$, where $\mathbf{x_i}$ are the support vectors, y_i their labels, α_i their weights (found by quadratic programming), and $K(.,.)$ is the kernel function. This distance is the output of the SVM classifier.

SVMs are capable of coping with high-dimensional and sparse features, which makes them ideal for classifying text [Joachims98] or high-level speaker recognition features, which have similar attributes.

2.3.4 Joint Factor Analysis

A recent and successful trend in tackling mismatch between training and test material (for example, if different voice capturing devices are used), is Joint Factor Analysis (JFA) [Kenny04, Kenny06, Kenny08, Burget09, Scheffer10]. It is based on the GMM-UBM approach described in Section 2.3.2, where each speaker is represented by a Gaussian Mixture Model which is derived by MAP adaptation from a universal background model. For JFA, the speaker model is transformed into a so-called *supervector* by concatenating the mean vectors of the Gaussian mixtures. Thus, every speaker can be represented by a FK-dimensional supervector, where F is the feature dimensionality and K the number of Gaussian mixtures. However, if different GMM models for the same speaker are trained with data from different acoustic conditions or channels, the resulting models and supervectors will also differ. This unwanted variance can be addressed by Joint Factor Analysis: each supervector is modelled as a linear combination of channel-dependent components and speaker-dependent components, disentangling the characteristics of speaker and channel. The idea behind this is that the speaker- and the channel-dependent compo-

nents can be described as points in lower-dimensional subspaces of the supervector space: the speaker subspace and the channel subspace. In order to do this, hyperparameters describing the channel and speaker spaces have to be estimated during system set-up. Then, a supervector M from one training sample of a speaker can be decomposed into two statistically independent parts, one stemming from the speaker, s, and one from the channel, c:

$$M = s + c. \tag{2.7}$$

The channel component c is modelled in channel space:

$$c = Ux, \tag{2.8}$$

where U is the *eigenchannel matrix*, describing the channel space, and x is the *channel factor* vector, describing the training sample's point in channel space. The columns of U are the eigenchannels, the eigenvectors which span the channel space, of which typical systems (like in [Kenny08]) have 50 to 100.

In a similar fashion, the speaker component s is expressed as:

$$s = m + Vy + Dz. \tag{2.9}$$

Here, m is the supervector of the universal background model from which the speaker model was derived. V is the *eigenvoice matrix* describing the speaker space, and y is the *speaker factor* vector. V's columns are the eigenvoices, the eigenvalues spanning the speaker space, of which typical systems (like [Kenny08]) have 300. The last part, Dz, describes additional speaker variance that is not captured by the eigenvoices. However, Dz is only useful in cases where large amounts of training data per speaker (15–20 minutes) are available [Kenny08]. In other cases, D can be left unused and set to 0, like in [Brümmer10].

Matrices U and V (and D, if used) are the hyperparameters of the JFA model, which are estimated in the system set-up phase from large datasets with an approach similar to Probabilistic Principal Components Analysis (see [Kenny06, Kenny08]). These datasets must contain a large number of speakers (at least as many as the desired number of eigenvoices), with multiple recordings from the same speaker over different channels. After the system set-up, speaker models can be trained: y and x are jointly estimated for each training sample by deriving their posterior distribution from statistics calculated from the training sample (see [Kenny06, Kenny08]). As x represents the training sample's channel, it can be discarded after training. Only the speaker information captured in y, which describes the speaker's characteristics, is retained as the speaker model and used for recognition.

2.4 Normalisation and Fusion

In addition to optimising features and speaker classification, commonly used strategies for improving speaker recognition performance and robustness are *feature and score normalisation* and *fusion* of information from different systems.

Normalisation can be applied at various stages in the recognition process. Often, feature and score normalisation are employed, after feature extraction and speaker scoring respectively. The most commonly used approaches used for this will be described in the next sections. Normalisation for the speaker modelling stage has also been investigated; currently, Joint Factor Analysis is a common approach to model normalisation. It is used to separate the characteristics of speaker and channel in a recording and can be regarded as channel normalisation for the speaker models (see the previous section).

2.4.1 Feature Normalisation

Feature normalisation aims at making the features more robust against disturbances — such as by noise or channel effects — so that they mirror the characteristics of the speakers and not of the recording conditions. The simplest approach to this is well known from speech recognition: the subtraction of the mean feature vector from each feature vector of an utterance, called cepstral mean subtraction (CMS) or cepstral mean normalisation (CMN) in the case of MFCC or LPCC features. This removes constant noise and channel effects present during the whole utterance, but it can also destroy speaker information. It seems to hurt performance on clean data with matching conditions but can improve recognition rates in the case of channel mismatch [Pelecanos01].

Mean normalisation can be extended by variance normalisation, like in [Viikki98]: a sliding window is shifted over the feature stream and the components of the feature vector in the centre of the window are normalised to have zero mean and unit variance, by subtracting the mean and dividing by the standard deviation of the window. Due to the sliding window, this method also addresses noise that changes over time. The variance normalisation counters the effect that additive noise tends to reduce the variance of the cepstral feature parameters [Pelecanos01].

More sophisticated adaptions of the feature vectors are *feature warping* [Pelecanos01] and *short-time gaussianisation* [Xiang02], which are frequently used in current NIST evaluation systems. There, a sliding window is shifted over the feature vector stream, and the streams for each vector component (cepstral coefficient) are considered separately. The window's central vector's components are mapped onto warped values (by using a look-up-table) so that the vector component streams conform to a normal distribution within the window. This counters a reduction in variance through additive noise and possible bias in the feature vector distribution due to channel effects, and it results in better speaker recognition

performance.

These and other feature normalisation methods and their usefulness for a NIST evaluation speaker recognition system are described in more detail in [Burget07].

2.4.2 Score normalisation

Score normalisation is used to map scores from different speakers to a similar range, so that they are easier to compare and so that a universal score threshold for speaker rejection can be set. The underlying idea is to use a set of non-target speaker scores (impostor scores) to derive a score mean μ_I and standard deviation σ_I [Auckenthaler00]. Those parameters are then used to normalise speaker score s:

$$s' = \frac{s - \mu_I}{\sigma_I}. \tag{2.10}$$

Depending on how the impostor scores are produced, different normalisation variants exist: For *zero normalisation (Z-norm)*, the speaker's model is tested on a number of example impostor utterances and from the scores of those utterances a speaker specific mean and variance is derived. These parameters are kept for future normalisation of all scores produced by the specific model.

While for Z-norm the utterances are varied and the model stays the same, for *test normalisation (T-norm)* it is done vice versa: For a given test utterance, scores are computed for a set of impostor speaker models. Mean and standard deviation derived from these scores can then be used to normalise all speaker scores produced for the test utterance.

The generation of Z-norm parameters can be done during the model training phase, while T-norm parameters must be calculated in the recognition phase, as this requires the test utterance. Both can be successfully combined and are usually used together, called ZT-norm. However, this method can fail if T-norm speakers and Z-norm test utterances don't match the application domain, that is, if the acoustic and channel conditions are different from the typical training and test material used. [Kinnunen10]

2.4.3 Fusion

Fusion of information from different speaker cues can dramatically improve recognition performance [Reynolds03b, Campbell03a, Reynolds03c]. Often, fusion is done at the score level, combining the scores of different speaker recognition subsystems, which may be based on different features or modelling approaches.

The simplest way to fuse N systems is by computing a weighted sum of their scores $s = w_0 + \sum_{i=1}^{N} s_i w_i$ (with $\sum w_i = 1$) for each speaker hypothesis. The weights can be set to equal values or adjusted experimentally. In [Brümmer06, Brümmer07], weights are optimised with linear logistic regression. For

this purpose, development data, for which the true speaker is known, is set aside to tune the weights on. Using scores produced for the training data is technically possible, however, they would likely give an over-optimistic impression of the systems' performance. For all subsystems to be fused, scores for the development data are produced with true speaker and impostor models. As the scores for a true speaker should be positive and scores for impostors negative, the system weights are then optimised so that the fused score is high for the true speaker tests and low for the impostor tests. This approach is frequently used in the literature and in NIST evaluations, a Matlab implementation can be found online[7].

Alternatively, the score values for an utterance stemming from the N subsystems can be regarded as an N-dimensional feature vector and fed into a classifier to yield a decision on the speaker hypothesis. In [Campbell03a], for example, several methods were tested: Perceptron, Radial Basis Function, Gaussian, Gaussian mixture, K-nearest Neighbour, Binary Tree, and Support Vector Machine classifiers were tried, and it was found that fusion with a GMM or perceptron yielded the best results in this case. The authors preferred the perceptron on the basis that it is simpler and yields decision boundaries that appear more reasonable. This suggests that a linear fusion method is at least comparable in performance to more complex approaches, which is supported by the results obtained in [Ferrer08] where two flavours of linear logistic regression are favourably compared with support vector machines and neural networks.

In [Ferrer08] an extension of linear logistic regression – which makes use of side information about the utterances – is proposed. The side information used in the paper are the utterance's the number of phones(obtained from an automatic speech recognition system), average of ASR word recognition confidences, and non-nativeness score. The non-nativeness score is a value obtained from an SVM classifier trained to categorise native and non-native speakers ([Shriberg08]). These continuous values are binned to obtain categorical features, so that each utterance can be assigned a type based on the side information. The LLR objective function is modified to train fusion weights for each subsystem and utterance type. When fusing the test utterances' scores, the weight of each classifier is adjusted depending on the test utterance type. This is based on the idea that different subsystems in the fusion may be better equipped to deal with different types of utterances. The papers shows that if suitable side information features are chosen and enough LLR weight training data exists, this approach benefits speaker recognition performance.

2.5 Summary

An overview was given over the field of speaker recognition and the research mainstream, which is in many cases driven by the biannual NIST Speaker Recognition Evaluations and the corresponding

[7] `http://sites.google.com/site/nikobrummer/focal`, last accessed: 2013-02-25

benchmark data. The commonly used performance measures, equal error rates and detection error trade-off curves, which are also to be used in this thesis, were described. The existing cues and features for speaker recognition were presented, with a focus on the various high-level features, which represent speaker information beyond the speaker's voice characteristics captured with spectral features. There are features for prosody, which are mainly based on the progression of the pitch, for pronunciation, which make use of phone and phone n-gram frequencies, and idiolect, which employ word n-gram frequencies. In broadcast data, it is also possible to spot a speaker's name with linguistic patterns and use this information in identification. A brief excursion to the field of authorship attribution for written text found that often word frequencies of topic-independent function words are used as features, but that topic information is very useful, if a way to extract it can be found.

The most popular methods for speaker modelling and classification were described, starting with simple log-likelihood classifiers. The standard speaker recognition baseline was sketched, Reynolds's GMM-UBM system, which is based on Gaussian Mixture Models. Support Vector Machines were delineated as they are often used for speaker recognition with high-level features because they are discriminative classifiers and because of their capability to cope with high-dimensional features. A brief introduction into the currently popular Joint Factor Analysis was given, which separates speaker characteristics from channel characteristics. Finally, methods for feature and score normalisation and for the score fusion of different speaker recognition systems were described.

Chapter 3

Topic-based Speaker Recognition

3.1 Motivation

As can be seen from the previous chapter, current speaker recognition systems rely on a number of speaker cues, including voice characteristics, prosody, and idiolect. However, the context of the speaker utterances is seldom taken into account and hard to model. One step towards using context for speaker identification is to infer likely speakers for an utterance based on their expertise for the topic at hand. In appropriate domains, such as broadcast news, the speakers can be expected to talk about their interests or fields of expertise. A tennis player will mostly talk about sports while a minister will likely speak about politics. Vice versa, if an utterance is about sports, it is more likely to come from the tennis player than from the politician. For authorship attribution in written text, a similar effect was found in personal communication [Goldstein-Stewart09]. Thus, topic information can be used in a speaker identification system, if a way to automatically extract it can be found.

In this chapter, two topic-based speaker recognition methods are proposed. Both were developed considering the practical constraints explained in Section 3.2, which stem from the lack of perfect transcripts and labelled topic training utterances. The first step towards topic extraction, which is used in both methods, is the application of automatic speech recognition to the audio recordings, in order to obtain a word transcript. The topic information in the transcript is then condensed by text pre-processing, described in Section 3.3. The first proposed method, described in Section 3.4, models topics implicitly by learning the speakers' characteristic keywords, which will in many cases be the topic marker words of their preferred topics. The second method models topics explicitly with an unsupervisedly trained probabilistic topic model, which automatically finds domain specific topics in an unlabelled training corpus. The topic model is then used to generate topic assignment vectors for the utterances, which can then be

used as feature vectors in topic-model-based speaker recognition, detailed in Section 3.5.

3.2 Constraints for Topic Information Extraction

There are a number of requirements and constraints for the topic information extraction methods to be used for topic-based speaker recognition. They chiefly stem from the fact that in most realistic applications, perfectly annotated audio material in large quantities is not available for training and speaker recognition and has to be approximated by other sources.

The first constraint is that manual transcriptions will not be available for all audio data in most cases. Yet, as an utterance's topic information is mostly contained in its words, a word transcript has to be obtained. It is an obvious choice to replace manual transcriptions with the output of a large vocabulary speech recogniser (LVCSR), which greatly reduces the transcription cost. In many spoken document retrieval systems, LVCSR will be applied in any case to facilitate text search in the audio data. In this case the transcripts may be used for improving the speaker identification performance at very low extra cost. However, using LVCSR transcripts requires that the topic extraction methods are robust to word recognition errors and do not require syntactically or semantically correct sentences.

The second constraint is that manually labelled topic training data may also not be available for the desired domain. There are plenty of methods in the literature to automatically classify a text into topic categories with machine learning methods if (enough) manually labelled training data for those topics is available (see, for example, [Sebastiani02]). However, obtaining or annotating domain-specific labelled topic training data increases the cost for setting up a topic-based speaker recognition system for a new domain, possibly to the point where it is not feasible or worthwhile. In order to make topic-based speaker recognition usable with little effort in as many domains as possible, topic extraction methods that do not require pre-labelled topic training data are required. This indicates the use of unsupervised topic modelling methods.

Still, some data that defines the background will be needed: A large and representative collection of example utterances is used in the majority of speaker recognition systems to learn the average speaker traits. For topic-based speaker recognition, this entails, for example, the estimation of average word frequencies and the generation of topic models for the domain. As words span longer time frames than phones or spectral feature vectors (which are take from frames of 20-30 ms in duration), a speaker recognition approach based on words would need more audio background data than approaches based on shorter features. A sufficient amount of background audio data may very likely not be available, thus some sort of substitute is needed. Also, topic-based speaker recognition is domain dependent, thus a different background set may be required for each new domain – for example one background set for

politics and a different one for business meetings. In automatic speech recognition, when transcripts for training the language model are sparse, written text is often used instead, as large text collections are more readily available than large audio corpora, especially since the Internet offers a wide variety of text material to be easily retrieved. For topic-based speaker recognition, it reduces the cost of training a system for a new domain if written text is used as background data instead of audio transcripts. This adds the third constraint that the methods for topic extraction should be robust to the difference between written text, which is used as background material, and LVCSR transcriptions, which are used for speaker training and recognition.

In summary, in order to reduce system set-up and processing costs, both topic-based speaker recognition methods proposed in this thesis use automatic speech recognition transcripts, employ methods that do not require labelled topic training data, and use a background set of written text[1]. The methods have, in part, also been published in [Baum12].

Both methods make use of the topic information present in the words of an utterance, however, it is captured in different ways. The first models topics only implicitly, by learning speaker specific topic keywords. The use of topic keywords as features is adapted from [Baum09a]; in addition to the log-likelihood ratio classifier used there, Support Vector machines are explored as an alternative modelling approach. An advantage of this approach is that it requires no explicit topic models, and thus no topic training. But while the performance of this method strongly depends on the speakers' topic preferences, the topic information is mixed up with the speaker's idiolect and other speaker marker words, such as names and titles. It is not entirely clear, how big the influence of each of these factors is.

Therefore, a second method is proposed, which explicitly uses topics by employing unsupervised probabilistic topic models to map words to topic clusters. With the topic models, it is possible to generate topic histograms for training and test documents. The histograms are then used as features for topic-model-based speaker recognition.

The approaches themselves are described in this chapter while the implementation details are specified in Section 5.1.

3.3 Utterance Pre-processing

Both proposed systems have in common the pre-processing chain that is applied to all material used for training and testing in order to normalise it and emphasise the contained topic information. The pre-processing steps are summarised in Figure 3.1.

For the audio material, the first step is the transcription of the words in the utterance with an

[1] The background set used for the experiments in this thesis is described in Section 4.4.

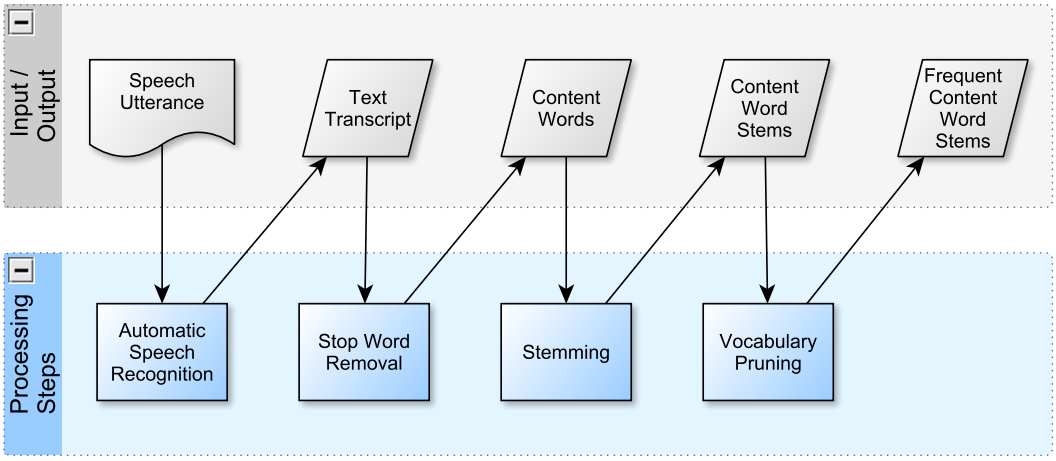


Figure 3.1: Utterance pre-processing for topic based speaker recognition.

automatic speech recognition system, yielding a text transcript. This transcript may contain syntactically or grammatically incorrect sentences, due to ASR recognition errors, and may lack proper punctuation. For the text documents from the background set, this step is unnecessary, obviously. In order to normalise the text documents, all punctuation and non-alphabetic characters are removed from them, as well.

Then, all stop words are removed from the utterances, as is commonly done in text categorisation systems. Stop words, also called function words, are the few hundred most frequently occurring words in a text collection (like “the”, “and”, “or”, ...), but they generally do not convey information about the topic, thus they are discarded. The remaining words are the ones that describe the thematic content of the utterance and will be referred to as “content words” in this text. As a side effect, this step reduces the difference between written text and ASR transcripts, as it removes words that are frequently misrecognised by ASR systems because they are short [Shinozaki01, Goldwater10].

In the next step, the remaining content words are reduced to their word stem, which is considered to contain the thematic information. This is commonly done in text categorisation and information retrieval to map different inflected variants (like “talk”, “talking”, “talks”, “talked”) onto one representation (“talk”). As the methods presented in this thesis are tested on German data, this step is particularly important because German is an inflecting language with a high number of variants per stem. The discarded affixes are mostly not relevant for topic identification. Also, this step removes a common type of ASR recognition error, as morphological substitutions, like “talks” instead of “talk”, frequently occur in ASR [Goldwater10]. This further reduces the difference between written text sources and ASR output. In experiments described in Section 5.5.2, stop word removal and stemming improved topic-based speaker recognition performance considerably.

Finally, infrequently occurring word stems are discarded. For this step, a dictionary of all stems occurring twice or more in the background and speaker training data must be built. Word stems not

contained in this dictionary are removed from the utterances. Discarding rare stems seems to have negligible effect on speaker recognition performance but greatly reduces the number of word stem types, making the speaker recognition systems more efficient.

3.4 Keyword-based Speaker Recognition

The first proposed topic-based speaker recognition approach makes use of topic or speaker marker keywords and therefore is called "keyword-based speaker recognition" in this thesis. It models topics only implicitly by learning the speakers' idiosyncratic keywords, which will be topic marker words for the speakers' preferred topics in many cases. An advantage of keyword-based speaker recognition is that it requires no explicit topic models, and thus no topic training.

The approach was initially explored in [Baum09a], where speaker-specific keywords were tested as features for recognising politicians by their fields of expertise, using log-likelihood ratio scoring. This speaker scoring method is described in Section 3.4.1. In addition, an alternative speaker modelling method is described in Section 3.4.2: Speaker classification with Support Vector Machines trained on tf-idf vectors. An experimental comparison between both modelling methods is given later in Section 5.5.1.

3.4.1 Log-likelihood Ratio Scoring

The log-likelihood ratio scoring method used for topic-based speaker recognition in [Baum09a] is based on similar approaches used in [Andrews02], [Doddington01], and [Baker04] for idiolectal and phonetic speaker recognition. There, phone or word n-gram frequencies in the speakers' utterances are compared to the same n-grams' frequencies in a background model (which represents the average speaker) in order to capture speaker-specific pronunciation or idiolect. Thus, speaker-specific word frequencies in ASR transcriptions compared with a background word frequency model seem a natural choice for capturing the speakers' favourite topics indirectly through keyword frequencies.

Figure 3.2 gives an overview of the method. All input into the system is preprocessed as described in Section 3.3. The background and speaker model training and the scoring done in the recognition phase are explained in the next two sections, respectively.

Feature Extraction and Speaker Modelling

The features in the log-likelihood ratio scoring method are word stem type[2] occurrence counts over the speakers' utterances and over the background (see Section 2.3.1 for an introduction to the use of log-

[2]Note that in this text "type" refers to the abstract word, like a dictionary entry, while "token" is the occurrence of the word in a text. Thus in the text "Yellow and orange and green" there are 4 types (yellow, orange, green, and) but 5 tokens (2 tokens for type "and").

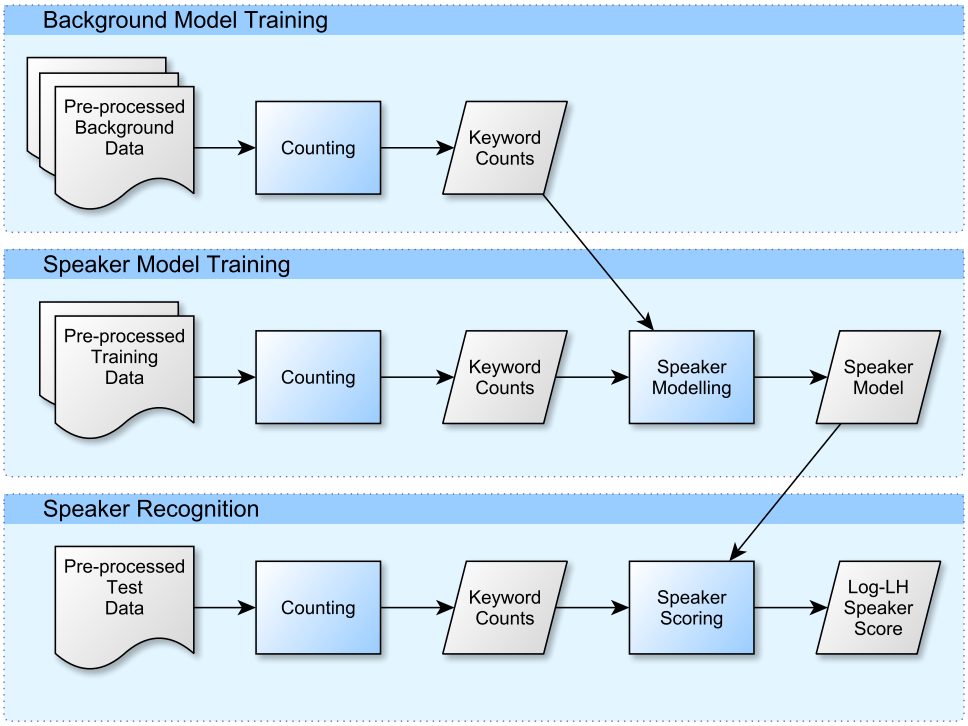


Figure 3.2: Overview for the log-likelihood ratio scoring implementation of the keyword-based speaker recognition system.

likelihood ratios in speaker recognition). In training, they are used to see which words a speaker uses more (or less) frequently than the average population. During recognition, the occurrence of a speaker's favourite words is taken as evidence towards this speaker being the true speaker, while words infrequently used by the speaker are taken as counter-evidence. Thus, the feature extraction consists of counting how often each word in the dictionary occurs in the given input material. This yields word frequencies for the average population in the case of the background data and speaker-specific word frequencies in the case of speaker training data. Each word's frequency in the speaker data can then be related to its frequency in the background, giving a measure $\lambda_i(n)$ of how specific word n is for speaker i:

$$\lambda_i(n) = log\left[\frac{H_i(n)}{N_i}\right] - log\left[\frac{H_{BG}(n)}{N_{BG}}\right], \tag{3.1}$$

where $H_i(n)$ and $H_{BG}(n)$ are the occurrence counts of word n in speaker i's training material and the background material, while N_i and N_{BG} are the total number of words in the speaker training data and the background, respectively.

As it was seen in [Baker04] that *maximum a posteriori* adaptation can halve the training data requirements for phonetic speaker recognition, MAP-adaptation is also used here. Analogous to the approach taken in [Reynolds00], speaker word counts H_i for speaker i are combined with background word counts H_{BG} to derive new smoothed speaker word counts $\hat{H}_i$. The influence of the speaker training

data depends on the adaptation coefficient α, with high emphasis on the speaker data when $\alpha \to 1$ and high emphasis on the background data when $\alpha \to 0$. The optimal value for a speaker recognition system's α has to be empirically determined (Section 5.1.2 details the value chosen for the experiments in this thesis). The MAP-adapted speaker count for word n is calculated as follows:

$$\hat{H}_i(n) = \alpha \cdot H_i(n) + (1 - \alpha) \cdot H_{BG}(n) \tag{3.2}$$

Using the adapted speaker word counts $\hat{H}_i(n)$ in Equation 3.1 yields:

$$\lambda'_i(n) = log\left[\frac{\hat{H}_i(n)}{\hat{N}_i}\right] - log\left[\frac{H_{BG}(n)}{N_{BG}}\right], \tag{3.3}$$

With Equations 3.2 and 3.3, a speaker's model can be trained by calculating and storing a log-likelihood ratio for each word in the dictionary.

Scoring

In order to derive a speaker score s_i for speaker i on a test utterance, a weighted sum of the log-likelihood ratios $\lambda'_i(n)$ for all the words in the utterance is computed; analogous to [Andrews02], the words are weighted according to their number of occurrences:

$$s_i = \frac{\sum_n w(n)\lambda'_i(n)}{\sum_n w(n)}, \tag{3.4}$$

where the weighting factor $w(n)$ is determined by the word count in the test data $c(n)$ and the discounting factor d:

$$w(n) = c(n)^{1-d} \tag{3.5}$$

The discounting factor determines the influence of the word count, with equal weights for all words regardless of their number of occurrences when $d = 1$ and weighting proportional to the number of occurrences when $d = 0$. The optimal discount factor for a system has to be empirically determined (Section 5.1.2 details the value chosen for the experiments in this thesis).

The speaker score s_i is a weighted average of log likelihood ratios, thus it indicates that a speaker is more likely to have said the test utterance than the average person if $s_i > 0$ and vice versa. The scores from all available speaker models for the test utterance combined give the utterances' score vector.

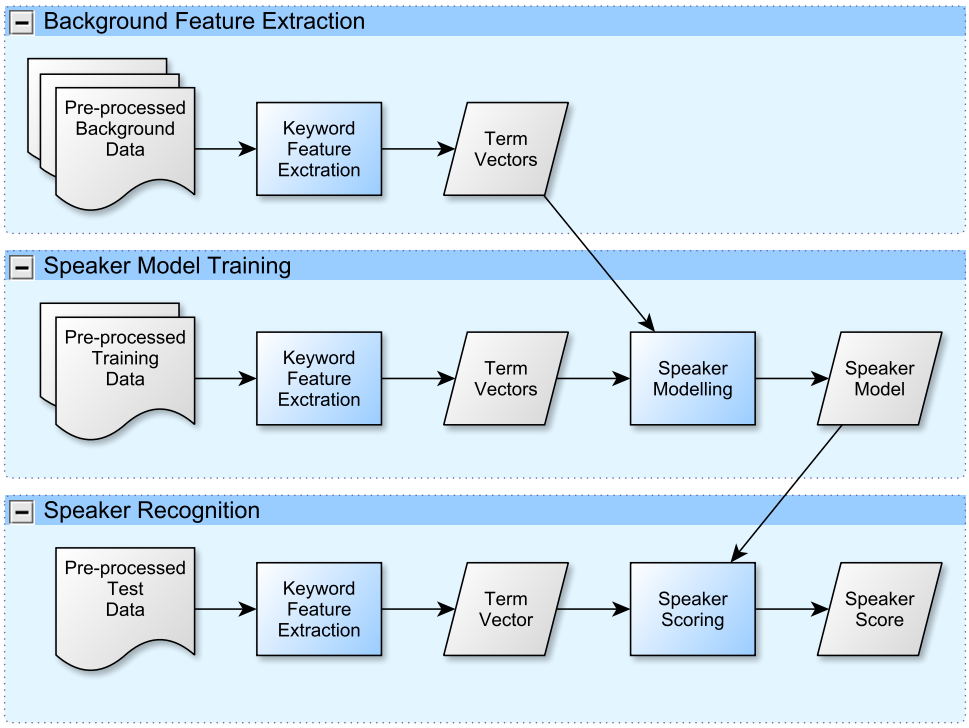


Figure 3.3: Overview for the SVM implementation of the keyword-based speaker recognition system.

3.4.2 Support Vector Machine Classifier

The second implementation of keyword-based speaker recognition uses weighted word frequency vectors as features for speaker modelling by Support Vector Machines (SVMs). It is based on text categorisation by SVMs [Joachims98], where the categories are speakers. This is similar to the authorship attribution in written text approach in [Diederich03]. The modelling is similar to the word-n-gram SVM speaker recognition system described in [Shriberg05], with the difference that the aim is to capture topic cues instead of idiolect, thus word stems are used instead of word n-grams.

Figure 3.3 presents an overview of the SVM implementation keyword-based speaker recognition. Before feature extraction, the content words are extracted as described in Section 3.3. The feature extraction method, modelling, and scoring are described in the next three sections, respectively.

Feature Extraction

The features used for keyword-based speaker recognition are weighted word stem frequencies, calculated from the utterances after the pre-processing chain described in Section 3.3 has been applied. Their extraction is depicted in Figure 3.4: The occurrences of each content word stem type in the utterance are counted, and only the counts are retained as the representation of the document, discarding information about word order. This simplification is common in text categorisation. Then, each utterance

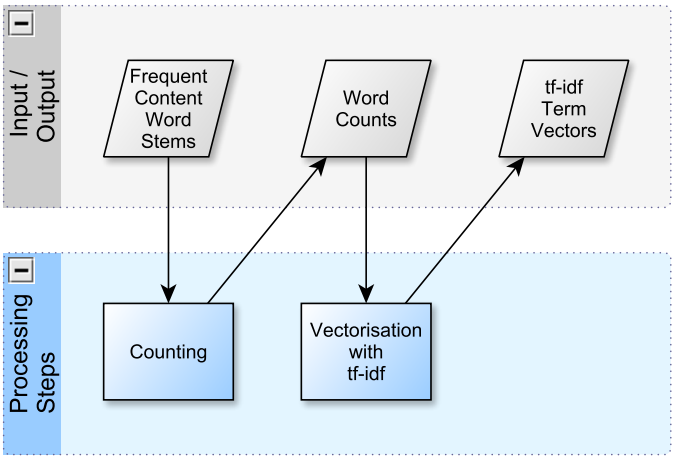


Figure 3.4: Tf-idf feature extraction for SVM keyword-based speaker recognition.

u is transformed into a term vector v_u with the term frequency — inverse document frequency (tf-idf) weighting. Tf-idf is well-known in text categorisation and information retrieval and yields good results for authorship attribution [Diederich03]. For the keyword-based speaker recognition system described here, the inverse document frequency is calculated on the background text collection:

$$v_u(w) = \frac{n_{w,u}}{\sum_k n_{k,u}} \cdot \log \frac{|D_{BG}|}{1 + |\{u_{BG} : w \in u_{BG}\}|} \tag{3.6}$$

The term vectors contain one entry per word stem type w (which is 0 if w doesn't occur in u). The number of occurrences of word stem w in utterance u is denoted by $n_{w,u}$, $|D_{BG}|$ is the number of documents in the background collection, and $|\{u_{BG} : w \in u_{BG}\}|$ is the number of documents in the background collection containing w. Thus, a word stem conveys important information about the utterance if it occurs frequently in an utterance and rarely in the background collection. The term vectors are N-dimensional, where N is the number of different word types in the dictionary. Also, they normally are very sparse, with non-zero entries for only a small number of dimensions.

An alternative term weighting method, rank normalisation, is also considered, because it is used in the idiolectal system described in [Shriberg05]. For rank normalisation, the word stem counts in all utterances in the background set are calculated, and a table A is generated that records in how many background utterances a word stem w occurs exactly x times, where table entry $A(w, x)$ is given by:

$$A(w, x) = |\{u_{BG} : n_{w,u_{BG}} = x\}| \tag{3.7}$$

For weighting, each word type count from an utterance is then replaced by its rank $r_u(w)$: the number of background documents where the term occurs less frequently than in the utterance. The rank is divided by the number of background documents $|D_{BG}|$ to normalise for background set size, yielding

a value between 0 and 1.

$$r_u(w) = \frac{\sum_{k=0}^{n_{w,u}-1} A(w,k)}{|D_{BG}|} \tag{3.8}$$

However, in experiments described in Section 5.5.3 standard tf-idf term weighting yielded better results for keyword-based speaker recognition than rank normalisation.

Modelling

After feature extraction, a soft margin binary Support Vector Machine classifier is trained for each speaker (see Section 2.3.3 for a short introduction to SVMs). As SVMs are discriminative classifiers, each speaker model is specialised to distinguish its speaker's characteristics from the rest of the population. The decision function outputs a measure for the likelihood of its speaker talking in the test utterance. A linear SVM kernel is used because this makes training fast and is known to give good results for text categorisation [Dumais98], authorship attribution [Diederich03], and idiolectal speaker recognition [Shriberg05]. The vector representations of the speaker training documents are used as positive examples and the background utterance vectors as negative examples, which effectively yields a classifier that separates speaker utterances from "average" utterances.

As the background contains vastly more documents than there are training utterances per speaker, it is essential to give more weight to the positive examples during training. Otherwise the overwhelming majority of negative examples would dominate the shape of the decision boundary. As described in [Morik99], this can be achieved in a soft margin classifier by introducing separate cost factors C_p and C_n to assign different costs to false positive and false negative classifications. The term to be minimised must be changed from $\frac{1}{2}\mathbf{w}^2 + C\sum_{i=1}^{n}\xi_i$ for the normal soft margin classifier (compare Section 2.3.3) to $\frac{1}{2}\mathbf{w}^2 + C_p\sum_{i:y_i=1}\xi_i + C_n\sum_{j:y_j=-1}\xi_j$. For the experiments in this study, misclassifying a speaker instance is penalised more strongly than misclassifying background instances by the ratio r of background documents to speaker utterances, which means that $C_p = r \cdot C_n$. This should correct the bias stemming from the uneven distribution of positive and negative examples.

Scoring

For speaker recognition, the weighted term vector of the test utterance is classified with the speakers' SVMs. Each SVM's output is the value of the SVM decision function for the utterance, which is the distance of the utterance vector from the hyperplane separating speaker from background vectors. This value is higher if the utterance vector is similar to the speaker vectors and lower if it is similar to the

background vectors, thus it can be used as a speaker's score for the utterance. The scores from all hypothesised speakers are then combined into a speaker score vector, which can be used in fusion.

3.5 Topic-model-based Speaker Recognition

While the method described in the previous section heavily depends on topics as speaker cues — because speaker specific terms in broadcast data will frequently come from the field of expertise of the speaker — topics are not modelled explicitly. It is possible that the speakers are identified by favourite terms that convey no topic information, such as filler words or other idiolectal cues, just like intended for example in [Doddington01]. In order to see how much speaker information really is conveyed by the topic of an utterance a second approach was devised which assigns the utterances to sets of topics and only uses this assignment as information for speaker recognition. As this approach cannot make use of idiolectal cues, it may perform worse than the keyword-based approach in theory. However, the explicit topic assignment could also give a performance gain because the topic classifiers have more than just the speaker's training data as input.

The set of topics first has to be defined, as it obviously depends on the application domain of the speaker recognition system. It is desirable to do this automatically, to reduce the set-up cost for topic-based speaker recognition on a new domain. The choice of methods for topic detection is limited by the constraints given in Section 3.2, especially by the lack of labelled topic training data. The topics and their marker words are to be learnt from the (written text) background training data, as this set is taken to represent normal utterances from the domain, discussing all the relevant topics. This requires the use of an unsupervised topic clustering method which is also able to assign topics to new documents that were not contained in the training corpus. Also, as an utterance may contain several topics to varying degrees, it is desirable to use soft topic assignments, for example by giving the probability of each topic being present in the utterance.

In this study, Latent Dirichlet Allocation (LDA), described in [Blei03b, Griffiths04, Steyvers07], was chosen for topic modelling as it fulfils the requirements and is popular, well-tested, and efficient. From a training corpus it learns a generative model for domain documents, in which the topics are latent variables. No manually assigned topic labels for the training documents are necessary, only the number of topics has to be decided upon by the user.

Figure 3.5 gives an overview of the topic-model-based system. Just like for the keyword-based approach, the pre-processing chain from Section 3.3 was applied to extract the content words. The Latent Dirichlet Allocation approach from [Blei03b, Griffiths04, Steyvers07] is summarised in Section 3.5.1 for the derivation of the topic model and Section 3.5.2 for the topic mixture feature extraction process. The

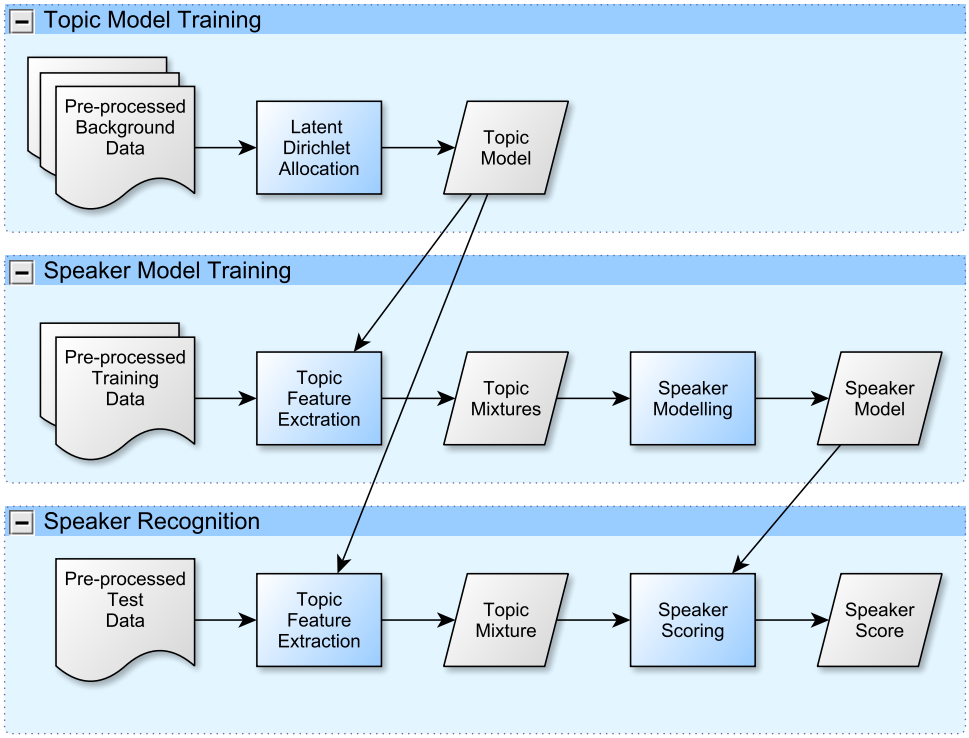


Figure 3.5: Overview for the topic-model-based speaker recognition system.

topic model was trained from the texts used as background material in the keyword-based approach. Three different possible speaker modelling methods and the corresponding scoring schemes are described in Sections 3.5.4 through 3.5.6.

3.5.1 Topic Derivation

Latent Dirichlet Allocation (LDA) [Blei03b, Griffiths04, Steyvers07] – which is used for topic derivation – learns a generative document model for a corpus in an unsupervised fashion. It assumes that each document contains a mixture of topics and each topic is linked to specific content keywords. The topics are the latent variables connecting documents and content words. The corpus is described formally as follows: it consists of D documents, has W unique word types, and it is assumed to contain T topics (the number of topics must be chosen by the user). Each word token i in the corpus has a word type index w_i, pointing to an entry in the corpus dictionary, and a document index d_i, giving the number of the document.

In probabilistic topic models, each topic j is a probability distribution $\phi^{(j)}$ over word types, describing how strongly each word type is associated with the topic and thus which words are important for the topic. Using these topic distributions, each document d can be represented as a mixture of topics, $\theta^{(d)}$, (where $\sum_j \theta_j^{(d)} = 1$). The principle is illustrated in Figure 3.6 with three topics whose words cover

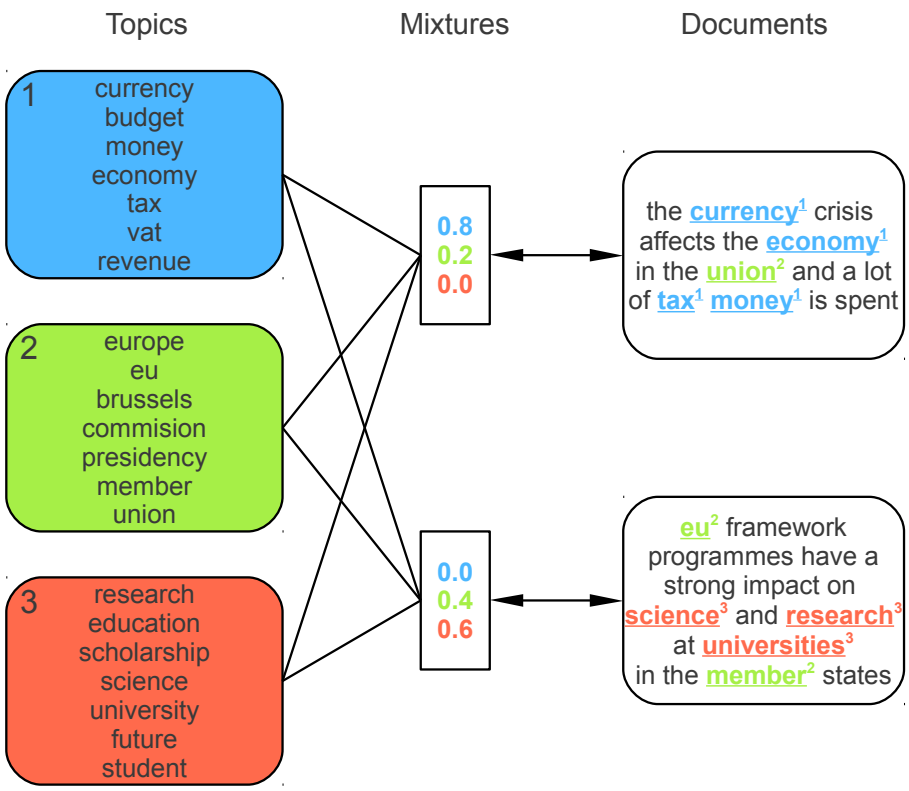


Figure 3.6: Representing documents as topic mixtures.

finance (topic 1, blue), the EU (topic 2, green), and science (topic 3, red). If the number of topics is much smaller than the number of word types, representing a document by its topic mixture can be thought of as a dimensionality reduction.

In LDA, the topics ϕ and topic mixtures θ of the domain corpus documents are approximated by Gibbs sampling, a Markov chain Monte Carlo algorithm [Griffiths04, Steyvers07]. Gibbs sampling is used instead of Expectation-Maximisation to avoid problems with local maxima of the likelihood function. It does not estimate θ and ϕ directly, rather each word token i in the corpus is assigned a topic j; the assignment is expressed as $z_i = j$. It is this assignment that is iteratively refined, progressing from random initialisation to (hopefully) meaningful association. (Note that two tokens of the same word type can be assigned different topics depending on the document in which they occur. For example in the model from Figure 3.6, the word "euro" might count towards the finance or the EU topic, depending on its context.) After the token-topic assignments are estimated, the type-topic and document-topic assignments, ϕ and θ, can be approximated from them.

During the Gibbs sampling estimation process, two matrices with counts are used: C^{WT} (a $W \times T$ matrix) contains the current assignments of word types to topics, and C^{DT} (a $D \times T$ matrix) contains the current assignments of documents to topics. The entries of the type-topic matrix C^{WT}_{wj} are the number of times word type index w is currently assigned to topic j. The entries in the document-topic matrix

C^{DT}_{dj} are the number of word tokens in document number d which are currently assigned to topic j. During one Gibbs sampling iteration, the count matrices are updated, considering each word token in the corpus in turn: First, both count matrices are decreased by 1 for the topic currently assigned to the token i in question (at positions $C^{WT}_{w_i z_i}$ and $C^{DT}_{d_i z_i}$). Then, the probabilities of assigning each topic to token i are estimated, conditioned on the topic assignments of all other tokens $\mathbf{z}_{other}$, the token's type index w_i and document index d_i, and all other factors such as the hyperparameters α and β according to Equation 3.9. According to this distribution, a new topic assignment z'_i for token i is randomly picked (sampled). (Sampling is used instead of just taking the most likely topic to avoid local maxima.) Finally, the entries in the matrices for the new token-topic-assignment z'_i, $C^{WT}_{w_i z'_i}$ and $C^{DT}_{d_i z'_i}$, are increased by 1, and the next word is considered.

$$P(z_i = j|\mathbf{z}_{other}, w_i, d_i) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \quad \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{3.9}$$

The first factor on the right hand side of Equation 3.9 corresponds to the probability of word type w_i if the topic is j, while the second factor expresses the probability of topic j in document d_i. Thus, the first factor "pulls" the topic assignment z'_i of word token i towards the topics which have its type w_i among their most probable words. The second factor "pulls" the assignment towards the topics of the document that i is from (regardless of the word type). During topic derivation, this causes word types that frequently occur together in the same documents to be assigned to the same topic.

The hyperparameters α and β determine the Dirichlet prior on the type-topic and document-topic distributions and can be interpreted as prior observation counts before any actual word tokens have been seen. They define the amount of smoothing on the distributions: Large hyperparameters lead to strong smoothing because a large number of tokens has to be observed to "overrule" the prior observation counts; the resulting distributions then are close to uniform. See Section 5.1.3 for the values of α and β used in the experiments in this thesis.

The steps in one Gibbs iteration can be summed up as follows:
For all tokens i in the corpus:

1. Decrease matrix entries $C^{WT}_{w_i z_i}$ and $C^{DT}_{d_i z_i}$ by 1 (remove influence of old assignment)

2. Compute the probability of token i belonging to each topic j, $P(z_i = j)$, with Equation 3.9.

3. From the computed probability distribution P, sample a new topic assignment z'_i for token i

4. Increase matrix entries $C^{WT}_{w_i z'_i}$ and $C^{DT}_{d_i z'_i}$ by one (record influence of new assignment)

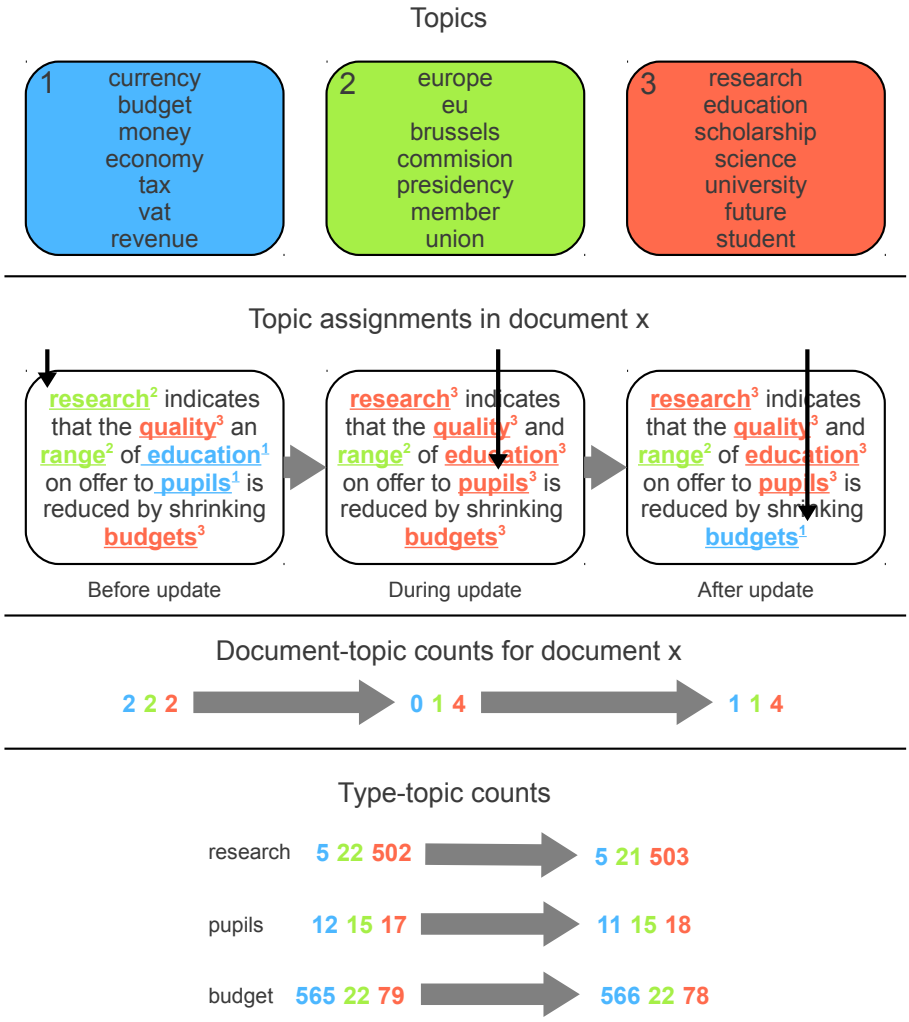


Figure 3.7: Updating the token-topic assignments in document x. The black arrow marks the token currently under consideration.

The required number of iterations is a parameter to be heuristically determined as checking for convergence in LDA with Gibbs sampling is still an unsolved problem.

As an example, consider a topic model with the 3 topics from Figure 3.6: During the current Gibbs iteration, the topics have converged somewhat towards finance for the first topic (blue), EU for the second topic (green), and science for the third topic (red) – because words from those areas frequently occurred together in the same documents in the part of the corpus that has been analysed so far. However, the Gibbs iteration is not yet completed, the token-topic assignments of other documents still need to be updated. Assume that document x – one of the remaining documents – consists of the sentence "Research indicates that the quality and range of education on offer to pupils is reduced by shrinking budgets." For an illustration of the example see Figure 3.7.

Before the update, the tokens in the document are still randomly assigned to topics from the ran-

domised initialisation. (For the sake of brevity, only nouns are considered in this example.) The entries in the document-topic count matrix for document x reflect this (2 counts for each topic). The document is not yet converging on a prevalent topic. However, the type-topic matrix entries for words like "research" and "budget" have already been drawn towards specific topics during the processing of the previous documents – let's say counts for "research" are 5 (topic 1), 22 (topic 2), and 502 (topic 3), while counts for "budget" are 565 (topic 1), 22 (topic 2), and 79 (topic 3).

During the update, each token is considered in turn – in the illustration the black arrow marks the current token. The first token is "research", which has been randomly assigned to topic 2. Step 1 of the update process removes this assignment from the count matrices. (The document-topic counts for x are now 2 (topic 1), 1 (topic 2), 2 (topic 3). The type-topic counts for "research" are now 5 (topic 1), 21 (topic 2), 502 (topic 3).) Step 2 computes the probability distribution of the token "research" in document x belonging to either topic according to Equation 3.9. Assuming that "research" is a strong indicator for topic 3, i.e. that it has a high count for topic 3 compared to the other words in the dictionary, the first factor in Equation 3.9 will be high for topic 3. If "research" is no strong indicator for the other topics, the first factor will "pull" the distribution towards topic 3. The second factor does not exert a strong pull towards any of the topics because the counts for topic 1 and 3 are equal. In step 3, a new topic assignment is sampled from the probability distribution. Because the probability is high for topic 3, assume that topic 3 is actually sampled. In step 4, the document-topic and type-topic matrices are updated accordingly, thus the document-topic counts for x are now 2 (topic 1), 1 (topic 2), 3 (topic 3), while the type-topic counts for "research" are now 5 (topic 1), 21 (topic 2), 503 (topic 3). After that, the next token is updated.

Assume that by chance the assignments for "quality" and "range" stay the same, and that education is assigned to topic 3 for the same reasons as "research". Now consider the update of token "pupils": A prevalent topic for document x has emerged; after Step 1 of the update of "pupils", the document-topic counts for x are 0 (topic 1), 1 (topic 2), and 4 (topic 3). "Pupils" is no strong indicator for any of the topics (yet) and has not been observed often in the corpus so far (type-topic counts for pupils are 12 (topic 1), 15 (topic 2), 17 (topic 3)). The probability distribution for "pupils" is now pulled towards topic 3 because it is the prevalent topic of x, so assume that "pupils" is now assigned to topic 3. If, during the processing of the remaining documents of the corpus, the type "pupils" is frequently found in documents with a prevalence for topic 3, it may become an indicator for topic 3. This in turn would pull documents containing this type towards topic 3.

The last word to be updated is "budgets". The prevalent topic of the document is topic 3, while the type is a strong indicator for topic 1, thus the probability distribution is pulled in two directions. Assume that topic 1 is sampled for "budgets" because the influence of the first factor in Equation 3.9 is

stronger in this case. This concludes the processing of document x in this Gibbs iteration. In the next iteration, the token-topic assignments may be refined.

When a sufficient number of iterations has been done and the token-topic assignments are sufficiently refined, the first and second factor of Equation 3.9 can be used to obtain the approximations of the document topic mixtures, θ, and word distributions that represent topics, ϕ:

$$\phi_w^{(j)} = \frac{C_{wj}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta} \qquad \theta_j^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{t=1}^{T} C_{dt}^{DT} + T\alpha} \tag{3.10}$$

In order to derive topic mixtures for new documents from the same domain, the type-topic distribution ϕ and respectively the type-topic counts C_{wj}^{WT} are retained. As the mixtures represent soft topic-document assignments, they can be used as feature vectors compactly describing the thematic content of a document or utterance.

3.5.2 Feature Extraction

In essence, the topic model training described in the previous section is a set-up phase for the topic feature extraction process. After estimating the type-topic distribution ϕ, respectively the type-topic count matrix C_{wj}^{WT}, the training and test utterances of the speakers can be converted to feature vectors by estimating their document topic mixtures for the trained topics.

Topic mixture extraction is done individually for each new document d'; ϕ is fixed and not changed during the feature extraction, so as not to distort the trained domain topics. The topic mixture estimation is again accomplished by Gibbs sampling, estimating the token-topic assignments of the tokens in the new document in a process very similar to the one explained in the previous section. Instead of random initialisation, now each token i is assigned the topic j that is most probable for the corresponding type w_i according to the type-topic count matrix C_{wj}^{WT}. This assignment is used to initialise a new document-topic count "matrix", which really is a document-topic count vector for just the new document, $C_{d'j}^{1T}$. Then, the token-topic assignments z_i (for all i in d') are refined through sampling from Equation 3.9, as before. However, now the type-topic assignments C_{wj}^{WT} are left unchanged, so that the new document-topic count vector $C_{d'j}^{1T}$ is adapted to represent the new document in terms of the existing topics. In summary, the process is the same as described in the previous section, only the type-topic count matrix is not updated, because the topics should not be changed after the topic derivation phase. As before, a user-defined number of Gibbs iterations is done to refine the document-topic count vector. After that, a topic mixture $\theta^{(d')}$ for the new document d' can be approximated according to Equation 3.10. This is

the feature vector for the new document, describing its topic content.

3.5.3 Restrictions on Speaker Modelling with Topic Feature Vectors

A number of machine learning methods for modelling speakers with topic mixture feature vectors are conceivable. However, the performance of those methods is limited in practise by a common problem in high-level speaker recognition: lack of large amounts of training data. In order to extract high-level features, in most cases more speech has to be observed as is the case for low-level features, because high-level features occur in larger time-frames. This means that more training material is required than for low-level speaker recognition. In the corpus used in this study (see Section 4), there are "only" 5 training examples per speaker, from which 5 vectors are extracted. This means that some modelling methods are not likely to exhibit good performance, such as Gaussian Mixture Models. In this thesis, three modelling methods are explored, which are expected to yield good results even with little speaker training data: Modelling with speaker prototype vectors (see Section 3.5.4), modelling with lazy learning (see Section 3.5.5), and modelling with Support Vector Machines (see Section 3.5.6).

3.5.4 Modelling and Scoring with Speaker Prototype Vectors

In the first method, each speaker is represented by the centroid of the topic mixture vectors computed from the speaker's training documents: the mean vector is calculated from the training vectors and used as an approximation of the speaker's preferred topics. This averages the speaker's topic preferences over several utterances, giving equal weight to all utterances regardless of their length.

For deriving the speakers' score on a test utterance, a measure of similarity of their topic preferences to the utterance's topic mixture has to be calculated. As the topic mixtures are probability distributions, well-known statistical similarity measures like Bhattacharyya, Jensen-Shannon, or Kullback-Leibler divergence are a natural choice. In experiments described in Section 5.5.5, Bhattacharyya, Jensen-Shannon, symmetric Kullback-Leibler, and cosine distance were used as scoring functions. All resulting systems performed comparably, with Kullback-Leibler giving slightly better results, so it was picked as the scoring measure.

A speaker i's score s_i is thus calculated as the negative symmetric Kullback-Leibler divergence between the speaker's centroid μ_i and the utterance vector u:

$$s_i = -D_{KL}(u, \mu_i), \tag{3.11}$$

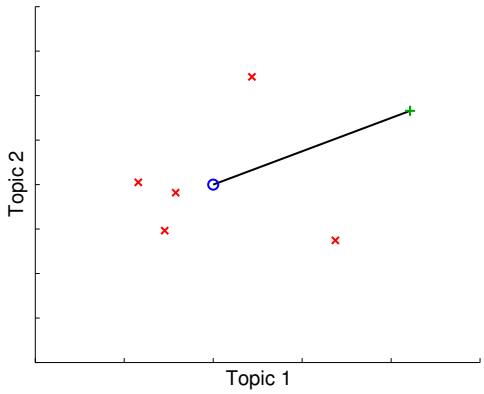


Figure 3.8: Scoring example with speaker prototype vectors.

where the Kullback-Leibler divergence between two distributions p and q is defined as:

$$D_{KL}(p,q) = \sum_{t=1}^{T} p(t) \log \frac{p(t)}{q(t)} + \sum_{t=1}^{T} q(t) \log \frac{q(t)}{p(t)} \tag{3.12}$$

Accordingly, the speaker with the highest score is the one whose centroid is the nearest neighbour of the utterance vector. The scores from all available speaker models for the test utterance combined give the utterances' score vector.

Figure 3.8 illustrates the modelling and scoring (in a 2-dimensional space, meaning that there are only two topics): from the training utterances (red x), the prototype vector (blue circle) is calculated. The score for the test example (green +) is the negative distance between prototype and test example.

3.5.5 Modelling and Scoring with Lazy Learning

The second modelling method is a lazy learner somewhat similar to nearest neighbour classification. Instead of deriving a speaker prototype from the training data, all training examples are retained to represent the speaker.

To derive a score for speaker i, the symmetric Kullback-Leibler divergences between the test utterance u and all speaker training utterances $t_{i1}, ..., t_{iK}$ are calculated and only the negative minimum is retained as the score.

$$s_i = - \min_{k \in \{1,...,k\}} D_{KL}(u, t_{ik}), \tag{3.13}$$

This scoring reflects the assumption that speakers can be interested in multiple topics, and that they are considered likely speakers for an utterance as soon as they have one similar training example, even if they normally prefer other topics. Figure 3.9 illustrates the principle: of all training utterances (red x), the one with the smallest distance to the test vector (green +) is identified (marked with a blue circle),

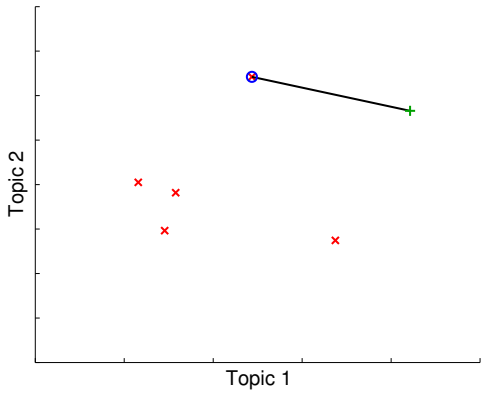


Figure 3.9: Scoring example with nearest neighbour lazy learner.

and the negative of this distance is used as the score.

Again, the scores from all available speaker models for the test utterance combined give the utterances' score vector.

3.5.6 Modelling and Scoring with Support Vector Machines

Support Vector Machines are used as the third modelling method, because they have been successfully applied in authorship attribution and speaker recognition with various features, even if the number of speaker training examples were small. The setup is very similar to the one described in Section 3.4.2 for the keyword-based approach. However, instead of the utterances' term vectors, their topic mixtures are used as the input for the SVMs. Again, one SVM classifier is trained for each speaker, using the speaker training utterances as positive examples and the documents from the background set as negative examples. During the recognition phase, the decision function output of each SVM on the test utterance vector is used as the score for the speaker it models, corresponding to the likelihood that the speaker talks in the test utterance. The decision function is the distance of the test utterance vector from the SVM's class-separating hyperplane. The scores from all available speaker models for the test utterance combined give the utterances' score vector.

As before, there are a lot more negative training examples (background documents) than positive ones (speaker training utterances). In soft margin SVM classifiers, this can be compensated for, as explained in Section 3.4.2: during training, the misclassification penalty for positive examples, C_p is greater than that for negative examples, C_n. In this study, C_p is larger than C_n by factor r, the ratio of background documents to speaker utterances, $C_p = r \cdot C_n$. This should correct the bias stemming from the uneven distribution of positive and negative examples.

The topic mixture feature vectors in the topic-model-based approach are much shorter than the term vectors in the keyword-based setup, and they have different properties – most importantly, they are

probability distributions. A different kernel function than in the keyword-based system is employed to make use of this additional information about the vectors' attributes: a Kullback-Leibler kernel. Moreno et al. proposed a Kullback-Leibler kernel for speaker recognition in [Moreno04b, Moreno04a]: A Gaussian Mixture Model is trained from the spectral features of each utterance and the Kullback-Leibler SVM is used to classify the utterances' GMM representation. Moreno et al.'s KL-kernel function is given in Equation 3.14, where $D_{KL}(u_x, u_y)$ is the symmetric Kullback-Leibler divergence. The scale and shift factors a and b were introduced "for numerical stability reasons" [Moreno04b, Moreno04a].

$$K_{KL}(u_x, u_y) = \exp(-a\ D_{KL}(u_x, u_y) + b) \tag{3.14}$$

This kernel is very similar to the standard Radial Basis Function kernel (see Equation 3.15), with the difference that symmetric Kullback-Leibler divergence is used as the distance measure instead of the Euclidean distance in the RBF kernel.

$$K_{RBF}(u_x, u_y) = \exp(-\gamma\ ||u_x - u_y||^2) \tag{3.15}$$

For the topic-based speaker recognition system in this study, the topic mixture vectors are used as utterance representations instead of GMMs of spectral features. The symmetric Kullback-Leibler divergence between topic mixtures is calculated with Equation 3.12.The parameters a and b have to be heuristically optimised for the task, as described in Section 5.1.3.

In order to assess the usefulness of the Kullback-Leibler kernel for the topic-model-based speaker recognition system, it is compared with the Radial Basis Function kernel in Section 5.5.6. The SVM modelling is also compared to the lazy learner and the speaker prototype approach there.

3.6 Summary

As topic information is useful for authorship attribution, and as context information in broadcast news can be expected to improve speaker recognition performance, this chapter introduced two approaches for topic-based speaker recognition and various ways of implementing them. The constraints for the realisation of topic-based speaker recognition in broadcast news were described: the lack of word transcriptions for audio data, the lack of labelled topic training data and the lack of sufficient background speaker audio data. Those problems are countered by the use of automatic speech recognition, of unsupervised topic learning, and of written text sources. This requires that topic based speaker recognition is robust to ASR transcription errors and the difference between transcriptions and written text. The difference between both is reduced via several pre-processing steps that densen the topic information of an utterance.

The first proposed approach, labelled keyword-based speaker recognition, models topics implicitly with the speakers' idiosyncratic words, which will mostly be topic marker words for their preferred topics. Two alternatives for implementing this with different feature extraction and speaker modelling are described. The first implementation employs log-likelihood ratio scoring: It relates the number of occurrences of each word in the speaker's training material with the number in the background set to arrive at a log-likelihood ratio for the word. This ratio captures how specific and idiosyncratic the word is for the speaker. During recognition, the ratios for the words in the test utterance are weighted and summed to arrive at a likelihood ratio score for the speaker. The second implementation works with word frequency vectors and Support Vector Machines for classification: The utterances are transformed into weighted word frequency vectors, either with tf-idf or rank normalisation. Then, a binary SVM classifier with linear kernel is trained for each speaker on the vectors for the speaker's training utterances. During recognition, the decision function output of the speaker's SVM is used as the speaker's score on the test utterance.

The second approach, labelled topic-model-based speaker recognition, aims to explicitly model the domain specific topics. First, the topics are automatically learned from the domain-specific background set with Latent Dirichlet Allocation. This does not require topic annotations for the training data as LDA is an unsupervised modelling method: It maps documents to topics and topics to words, using the fact that topic marker words for the same topic will frequently occur together in the same documents. With the obtained LDA topic model, each utterance can be transformed into a topic mixture vector, which represents soft assignments of the utterance to the set of domain topics. Using the topic mixtures as feature vectors, the speakers' topic preferences can then be modelled in different ways, three of which are explored here. The speakers' preferences can be represented by a prototype, an average of their training topic vectors, and their score for an utterance can be computed by taking the negative Kullback-Leibler divergence between the prototype and the utterances topic vector. This describes the speakers' main interests in the training data. Alternatively, a lazy learner can be used, which calculates the KL divergence between the utterance vector and each training vector of the speaker, using as the score the negative of the minimum distance. This puts more emphasis on side interests of the speakers, assuming that they are likely candidates for an utterance if they were observed talking about the same topic at least once. Finally, the use of Support Vector Machines with a Kullback-Leibler kernel is proposed, which uses the knowledge that topic mixture vectors are probability distributions. As SVMs have been successfully applied to speaker recognition, they may be also applicable to topic-model-based speaker recognition.

Overall, this means that for the two approaches the best-performing modelling method has to be selected from (a) log-likelihood scoring and (b) SVM-based classification for the keyword-based method

and (1) speaker vector prototypes, (2) lazy learning, and (3) SVM-based classification for the topic-model-based method.

Chapter 4

The Bundestag Corpus

As mentioned in Section 2.1.2, speaker recognition research is strongly driven by the benchmark corpora available for testing new algorithms. A lot of research is conducted and evaluated on the NIST SRE corpora, which mainly consist of telephone conversations between strangers of a few minutes length. In order to be able to compare new methods with the state of the art, it would be desirable to test on these collections. However, this poses a problem for topic-based speaker recognition methods for broadcast data for two reasons: First, the data is obviously not from the broadcast domain, so direct conclusions on the usefulness of the new methods for broadcast data would be difficult, although tendencies would be visible. Second, and more importantly, the benchmarks are combined from three main source corpora: Switchboard, Mixer, and Fisher [Cieri04b, Cieri04a, Doddington01]. These corpora predominantly contain telephone conversations between paid volunteers, recorded just for the purpose of obtaining speech data. Due to the recording protocols, the conversation topics were not chosen freely by the speakers: All conversations recorded on the same day were assigned the same topic, which was picked from a prepared topic list in order "to encourage a broad range of vocabulary" [Cieri04b]. Speakers were asked to talk about the assigned topic during the preparation phase of their conversation, however, "there was no penalty for conversations that strayed from the assigned topic" [Cieri04a]. Accordingly, it is unlikely that the speakers' topic preferences are clearly visible in the NIST benchmark and meaningful results for topic-based speaker recognition can not be expected on this data. Thus, to be able to test topic-based speaker recognition for broadcast data, a suitable corpus had to be found or created first. The resulting Bundestag corpus has also been described briefly in [Baum12].

4.1 Corpus Requirements

In order to be able to meaningfully test topic-based speaker recognition for broadcast data, the evaluation corpus has to fulfil a number of requirements:

1. The corpus should contain broadcast data, to allow direct conclusions on the usefulness of the new methods for the domain. This excludes a large number of existing corpora with telephone data, including the NIST SRE benchmarks.

2. The data should come from a realistic scenario in order to test the validity of the proposed approaches for real-life applications. Specifically, recorded data with prepared text prompts read by paid volunteers is undesirable, which further reduces the number of usable existing corpora. Rather, data on which a real-life search application would be built should be used.

3. Speakers should usually talk about their interests or fields of expertise, otherwise topic-based speaker recognition would not make sense. Of course, the corpus should not be manipulated towards fulfilling this condition, as this would distort any estimation of usefulness of the new methods. This relates to the previous requirement of realistic data. Again, this requirement excludes the NIST SRE benchmarks.

4. There should be a large number of recurring speakers in the data, ideally more than 100, to allow statistically meaningful analysis. For evaluation purposes, the material must obviously not only contain many speakers but their appearances in the data need to be annotated in some form as well. This excludes corpora where the speakers' identities are not annotated.

5. There should be enough material from each speaker to allow for training and testing. As idiolectal speaker recognition was tested with between 1 and 16 conversations (average conversation length 2.5 to 3 minutes) training material per speaker in [Doddington01], it was decided that there should be at least 5 training examples per speaker. There should also be an additional number of test samples to allow for a sufficient number of test trials (depending on the number of speakers).

Additional requirements for the purpose of this work were that the data should be in German, to be able to use the available German speech recognition system and to demonstrate the resulting search application to German customers, and that the data acquisition should not require a large budget.

As no such corpus was available from the European Language Resources Association (ELRA)[1] or the Linguistic Data Consortium (LDC)[2] at the time of the experiments, it had to be created.

[1] `http://catalog.elra.info/retd/index.php?cPath=37_46`, last accessed: 2013-02-25
[2] `http://www.ldc.upenn.edu/`, last accessed: 2013-02-25

A search for suitable material led to the German Parliament's "Web-TV" service[3] which offers video recordings of speeches given by politicians in German parliament. The videos are downloadable as high-quality MP4 files from 2009 onwards and fulfil all the requirements: The videos are produced for the parliament's internal TV station ("Parlamentsfernsehen") and can be licensed by other stations. Snippets extracted from these or very similar recordings appear in news broadcasts on a daily basis and constitute the relevant material that multimedia search engine users would look for. Also, the Web-TV archive already has a search engine which uses manually prepared metadata and would benefit from full spoken content search. Thus, the data is from the broadcast domain and is used in a real-life search application, fulfilling the first two requirements. Speakers in the videos are politicians who are very likely to talk about their fields of expertise, which, however, may change over time. Thus, the material allows for realistic evaluation of the usefulness of topic information in speaker identification, satisfying the third condition. There are approximately 600 members of parliament (due to German electoral law, the number of MPs in German parliament may vary between elections) and other politicians (ministers, secretaries of state, etc.), many of whom give speeches regularly. This ensures that for a large number of speakers sufficient training and testing material is available. Speeches can be retrieved from the Web-TV archive by speaker, making it possible to annotate the retrieved files with speaker identities. Thus, requirements 4 and 5 as well as the German language and cost conditions are met as well.

Note, that Bundestag data comes from only one "TV station", which probably yields more homogeneous material because of similar recording equipment and post-processing. This can be considered a limitation on the realism of the data. However, actual broadcast material still seems preferable to the NIST benchmark telephone data from the realism perspective.

4.2 Corpus Preparation

After the Bundestag recordings were selected as suitable material for experiments, the material had to be retrieved, checked, and an evaluation plan had to be defined in order to have a consistent testing procedure.

4.2.1 Consistency Checking

First, all available speeches from July 3rd 2009 to November 12th 2010 were downloaded as MP4 videos from the Web-TV website, and annotated with the date, time, and speaker information from the archive. The retrieved 4,090 videos were then submitted to a semi-automatic sanity and consistency check (as described below) to identify faulty MP4s and incorrectly annotated data: The videos were processed

[3] `http://webtv.bundestag.de/iptv/player/macros/bttv/index.html`, last accessed: 2013-02-25

with speech analysis tools, and files which had unusual output from these tools were manually inspected. This was necessary to reduce the cost of consistency checking to an acceptable amount, as a manual inspection of all videos would not have been feasible with the available budget.

The videos were transcribed with the ASR system described in Section 5.1.1, and 18 videos were removed because they were empty, incorrectly encoded (and thus could not be processed by the ASR), contained no audio or mostly noise or for other reasons had less than 50 words in the resulting ASR transcript. Furthermore, a preliminary spectral speaker recognition system (like the one described in Section 5.2.1) was set up and speaker models were trained from the available data for each speaker. The videos were then scored with the preliminary spectral system, and files which had an extremely low score for their true speaker were manually examined. The low score would indicate that the respective file's spectral content differed strongly from the average spectral content of the material available for that speaker. Of files with low scores, 7 had to be removed because they really contained an unknown speaker instead of the alleged politician, while another 17 files were manually cut because they included a substantial amount of speech from other speakers (due to long interjections, questions, and comments made by politicians in the auditorium). Note, that the preliminary speaker recognition system was only used for this purpose and that new models were trained for the evaluation described in Chapter 5 after the training and test sets had been selected.

As the consistency checks excluded 25 unusable files, 4,065 files remained for corpus construction ($4,090 - 18 - 7 = 4,065$).

4.2.2 Evaluation Protocol Design

After consistency checking, the remaining files were split into different sets for evaluation. It was decided that for a target speaker there should be at least 5 videos for training (which is the same order of magnitude as the 1–16 conversations training material per speaker in [Doddington01]) and 2 to 5 videos for testing. The number of test videos was selected in order to have as many target speaker trials in the evaluation as is feasible with the available data without giving too much weight to single speakers with many speeches (who might be especially easy or hard to recognise). Where further recordings were available for a speaker, between 2 and 5 of them were used as development material. Thus, politicians were selected as target speakers if there were at least 7 recordings of them in the retrieved data, which was the case for 235 speakers. For each target speaker, the chronologically first 5 speeches were selected as training material, and another 2 to 5 randomly (not chronologically) chosen speeches as test material. For the 120 speakers for whom there were more than 10 speeches, another 2 to 5 recordings were randomly selected from the yet unused speeches. Those were set aside as a development set, which could be used

to optimise system parameters without over-fitting to the test data. Table 4.1 sums up this evaluation design.

Number of target speakers	235
Number of target speakers with development material	120
Number of training files per speaker	5
Number of test files per speaker	2–5
Number of development files per speaker	2–5
Total number of training files	1,175
Total number of test files	994
Total number of development files	412
Number of target trials	994
Number of non-target trials	232,596

Table 4.1: Evaluation Design for the Bundestag corpus.

The resulting lists of training, development, and test speeches are available online at `http://www.ifs.tuwien.ac.at/ir/speakerrec/` and are attached to the electronic version of this document[4].

With this split into training, development, and test material, 245 videos from 45 target speakers remain unused for the evaluation because having many test trials from a small number of speakers may distort the obtained results if those speakers are especially hard or easy to recognise. Also, 1,222 recordings from 363 speakers for whom less than 7 speeches were online at the time of corpus construction are not used in the Bundestag corpus.

What is primarily measured in the evaluations in this thesis is the speaker verification performance, that is, the average performance of individual binary speaker classifiers. This is done because performance is independent of the target speaker set size, see Section 2.1.1. An important characteristic of a speaker verification corpus is the number of target and non-target trials. A target trial is a trial where the test file's score is computed with the model of the true speaker (target speaker) of that file. Tests with models from other speakers are non-target trials. This means that scoring one speaker's test video against another speaker's model counts as a non-target trial. As each speaker model is scored against all test videos, there are 232,596 non-target trials: 994 test files times 234 speakers who are not the true speaker equals 232,596 trials. The number of target trials is the number of test files, 994, as there is only one true speaker per test file.

[4]The attachments can be accessed through the following paper-clip icons: training , test , and development list

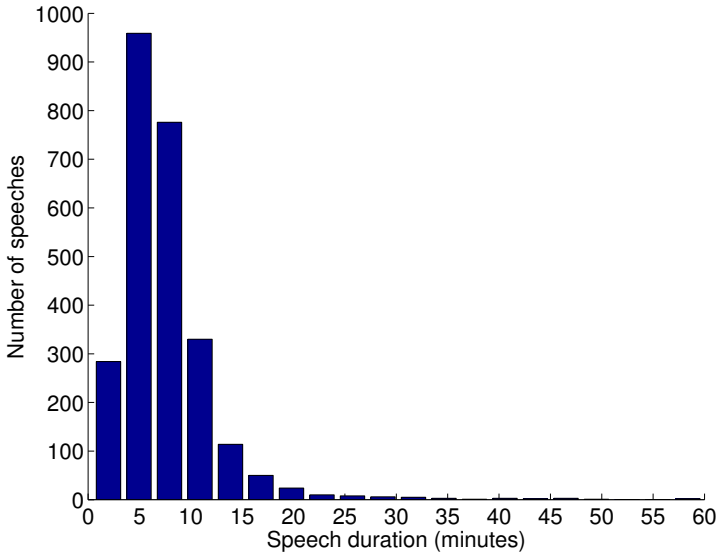


Figure 4.1: Histogram of video file durations in the Bundestag corpus.

4.3 Corpus Characteristics

After defining the train, test, and development sets for the Bundestag corpus, the contained files were inspected to get an impression of their characteristics.

Of the 235 speakers in the corpus, 71 (30%) are female and 164 (70%) are male, which corresponds to the overall proportion of women amongst members of parliament of 33%[5].

The training set amounts to a total of 1,175 recordings and approximately 150 hours of video, while the test set has 994 videos of about 120 hours overall length. There are 994 target trials and 232,596 non-target trials in the evaluation. Finally, the development set contains 412 videos with a total duration of approximately 54 hours. Overall, training, test, and development sets amount to about 324 hours of recordings of 2,581 speeches from the time span between the 3rd of July 2009 and the 12th of November 2010. The set sizes are also summed up in Table 4.2 for convenience. Automatic speech recognition was done on all videos, and the number of words in the resulting transcripts is also given for each set in Table 4.2.

Set	**Number of videos**	**Total duration**	**Number of running words**
Train	1,175	150 h	1,141,815
Test	994	120 h	927,444
Development	412	54 h	407,661
Total	2,581	324 h	2,476,920

Table 4.2: Train, test, and development set sizes in the Bundestag corpus.

[5] `http://www.bundestag.de/bundestag/abgeordnete17/mdb_zahlen/frauen_maenner.html`, in German, last accessed: 2013-02-25

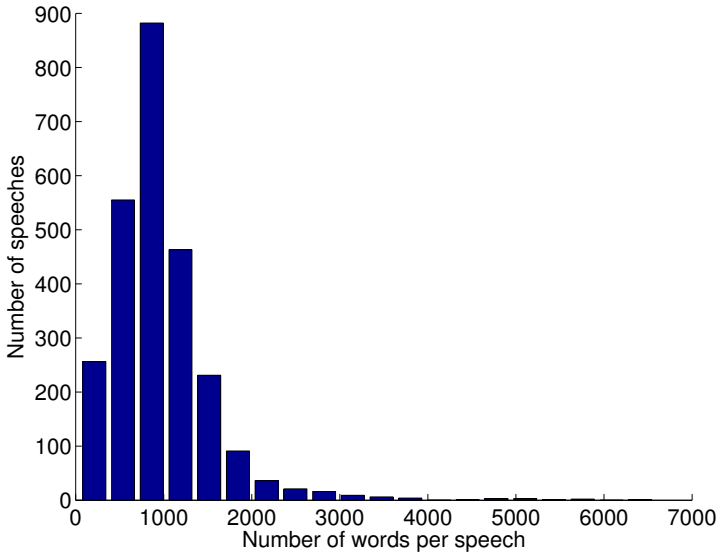


Figure 4.2: Histogram of number of running words per speech in the Bundestag corpus.

The length of the recordings ranges from 30 seconds (for a short question) to an hour (for the Chancellor's address to the newly elected parliament), but the majority of speeches is between 5 minutes (first quartile) and 9 minutes (third quartile) long, the median being 6:30 minutes (see also Figure 4.1). The median for the number of words recognised by the automatic speech recognition system per speech is 852, the first and third quartile are 643 and 1,176, respectively. Thus it can be expected that a typical speech is roughly between 600 and 1,200 words long (see also Figure 4.2).

Figure 4.3 gives an impression of the video material and of the environment in the Bundestag. The acoustic conditions found in the recordings are usually rather good: not studio quality but not heavily degraded either. The controversy of the speaker or discussed topic varies and with it the amount of emotional speech, high vocal effort (shouting), and background noise. Speakers are heckled from time to time, so there is crosstalk, and the microphones are not always optimally set-up, which occasionally leads to reverberation or acoustic feedback. The politicians are semi-professional to professional orators, and although speeches are not normally read, speakers often use prepared notes for guidance. Accordingly, the resulting speech is semi-spontaneous.

4.4 Text Data from the Bundestag Domain

As topic-based speaker recognition was to be evaluated with a written text background collection, text data from the domain had to be obtained. As expected, there is much more text than video material in the parliament's archive. The stenographic transcripts of all sessions of the Bundestag since 1998 are available

Figure 4.3: Example screenshots from the Bundestag videos (taken from Angela Merkel's speech on the 19th of May 2010, 9 a.m.).

in machine-readable PDF format[6]. These transcripts are manually produced by clerks; note that they are not always literal renderings of what was being said: The speakers have the right to read and change them (remove hesitations, correct grammatical errors, etc.), as long as the sense and meaning of the speech and its parts are not changed. A previous evaluation on a small number of example speeches indicated that the congruence between the stenographic transcripts and accurate literal transcriptions, measured by word accuracy, is between 65% and 90% [Biatov02]. The stenographic transcripts closely resemble written text — although they are renderings of oral speech, hesitations, disfluencies, grammatical errors, etc., are removed and sentences are "cleaned up". Also, they obviously do not contain ASR errors, and thus they can differ substantially from the ASR transcripts used to train and recognise speakers in the experiments. They serve as a test for the robustness of the proposed methods to this difference and for the validity of the assumption that written text can be used as background material.

Figure 4.4 shows an example for the content and formatting of the PDF documents. In order to be able to use them in the two text data sets described in the next sections, the contained text had to be extracted. The two column formatting, hyphenation, headers, footers and markers were removed. Then, the transcriptions were segmented into speeches by text pattern matching and saved as separate documents. This was done by manually inspecting randomly selected PDF documents and deriving a heuristic for recognising the beginning of a new speech: New speakers are in most cases marked in the text by their name, their position or role, and a colon, followed by a newline. Searching for these indicators, the text could be cut into documents (mostly) containing individual speeches. Statements from the presidents of parliament (Bundestagspräsidenten) were filtered out, as they are mainly concerned with the internal organisation of parliament and announcement of new speakers — and not the desired topics. This was done by removing speeches where the speaker's role was given as "Bundestagspräsident" or just "Präsident".

From this collection of text documents, two sets were extracted for the experiments: a speaker

[6] `http://www.bundestag.de/dokumente/protokolle/index.html`, last accessed: 2013-02-25

11478 Deutscher Bundestag – 16. Wahlperiode – 111. Sitzung. Berlin, Mittwoch, den 12. September 2007

Rainer Brüderle

(A) Frau Bundeskanzlerin, wann setzen Sie die Steuersenkung um, die auch Ihr Wirtschaftsminister fordert?

(Renate Künast [BÜNDNIS 90/DIE GRÜNEN]: Das hätten Sie doch gestern besprechen können!)

Sie reden davon, den Aufschwung stärken zu wollen. Wenn Sie entsprechend handeln wollen, dann sollte sich die Bundesregierung die Vorschläge zur Steuersenkung aus dem Wirtschaftsministerium zu Herzen nehmen und umsetzen. In diesem Fall bedeutet von Glos zu lernen, besser zu werden.

(Beifall bei der FDP)

Trotz sprudelnder Steuerquellen behauptet der Finanzminister, dass es keinen Spielraum für Steuersenkungen gibt. Dabei führen niedrigere Steuern zu mehr Investitionen, mehr Wachstum, mehr Arbeitsplätzen und mehr Konsum. All dies bringt auch mehr Steuereinnahmen.

Präsident Dr. Norbert Lammert: (C)

Das Wort hat nun die Bundeskanzlerin Frau Dr. Angela Merkel.

(Beifall bei der CDU/CSU und der SPD)

Dr. Angela Merkel, Bundeskanzlerin:

Herr Präsident! Meine Damen und Herren! Wie immer die Opposition es dreht und wendet: Deutschland hat wieder allen Grund zur Zuversicht.

(Beifall bei der CDU/CSU und der SPD)

Wir sind dabei, eine lange Durststrecke zu überwinden. Die Wirtschaft wächst so stark wie seit sechs Jahren nicht mehr. Seit Regierungsantritt ist die Zahl der **Arbeitslosen** um mehr als 1 Million gesunken. Das ist der tiefste Stand seit 1995, das heißt seit nunmehr zwölf Jahren.

(Beifall bei der CDU/CSU und der SPD)

Figure 4.4: Example for the format of the stenographic transcripts (taken from the transcript of the 111th session of the 16th legislative period on the 12th of September 2007, page 11,478).

recognition background data set and a text corpus for evaluating the shift of speaker topic preferences over time.

4.4.1 Speaker Recognition Background Set

For the topic-based speaker recognition background set, all available texts from October 18th 2005 to December 19th 2008 were pooled, amounting to approximately 10 million running words in 18,452 documents. The time period was selected to contain material not too far in the past so that it would have similar topics and topic marker words. Also, data from 2009 onwards was excluded so that the background texts do not coincide with the speeches used for speaker training and recognition. This ensures that the stenographic transcript of such a speech cannot be in the background training set.

4.4.2 Speaker Preference Shift Evaluation Corpus

To evaluate the impact of speaker topic preference shifts on topic-based speaker recognition, a second text corpus was prepared, covering the timespan from 1999 to 2010. With more than 10 years of speeches, an estimation of the degree of change should be possible. Of course, ideally this evaluation should be done on transcripts of audio material, but, unfortunately, Bundestag video recordings are not available from the Web-TV archive this far into the past. Again, written text documents were used instead, to at least get an estimate of the performance change.

For this evaluation only documents with more than 100 content words were used, excluding short

questions, comments, and remarks and focusing on normal speeches. The speakers for these documents were determined by taking their name from the beginning of the first line. Those speakers who were members of parliament from 1999 to 2010 and who had 3 or more speeches in 1999 were selected as the target speakers, 97 in total. The speeches from all other speakers were put into the background set. After the speaker selection, the speaker names at the beginning of texts were removed, because they are an artifact of the stenographic transcripts and not contained in the real speech. However, any names that were part of the actual speech were left as they were.

The target speakers' documents were sorted into 11 sets by year, from 1999 to 2010, randomly selecting at most 5 documents per speaker and year, so as to achieve a corpus more balanced between speakers. The number of target speaker speeches per year and of background speeches is listed in Table 4.3. The intention is to train the topic-based speaker recognition systems on the data from 1999 and then evaluate recognition performance for each of the following years separately to track changes.

Year	**Number of documents**
1999	397
2000	365
2001	369
2002	353
2003	397
2004	376
2005	309
2006	337
2007	316
2008	314
2009	292
2010	347
speaker total	4,369
background	33,571

Table 4.3: Number of documents per set.

4.5 Summary

In this chapter, the corpus which is used for the evaluation of the proposed methods is described. The requirements for benchmark data for testing topic-based speaker recognition for broadcast news were assessed: The data must stem from the broadcast domain and from a realistic application scenario. The speakers should normally talk about their preferred topics or fields of expertise, there should be many speakers (at least 100 is desirable) and ample material for each of them. It was found that no such corpus was available at the time of this writing. Thus, a corpus had to be created from suitable material. Video recordings of speeches from the German parliament (Bundestag) were chosen as suitable and

easily obtainable material. Videos were retrieved, checked for consistency, and an evaluation protocol for speaker recognition was designed for them. The resulting corpus contains 235 speakers, a total of 324 hours of videos, and is divided into a training, test, and development set. Additionally, two matching written text corpora from the same domain were extracted from stenographic transcripts: One as a background set for speaker recognition system training, and one as an evaluation corpus for speaker topic preference shift over time. The evaluation of the proposed methods is given in the next chapter.

Chapter 5

Experimental Evaluation

This chapter describes the experiments that were carried out to explore topic-based speaker recognition. The methods described in Chapter 3 were implemented and evaluated. For the implementations, various tools were employed whose set-up and parameters are given in Section 5.1, to ease reproducibility of the results. To be able to compare the performance of topic-based speaker recognition on the Bundestag corpus with existing methods, two baseline speaker recognition systems were set up — one based on idiolectal and one based on spectral features (see Section 5.2). To estimate the significance of differences in performance between the various systems, McNemar's significance test was used; the method is described in Section 5.4.

Section 5.5 verifies assumptions made and compares the different feature extraction and modelling methods, to find the best set-up for the topic-based speaker recognition systems. Section 5.6 then assess whether topic-based speaker recognition works, and whether it provides speaker information that is not captured by systems based on other cues. The baseline systems were fused with both topic-based speaker recognition systems to determine whether the topic information improves the overall performance. The processing chain for the combined systems is illustrated in Figure 5.1. Also, the influence of speaker interest changes on topic-based speaker recognition is investigated in Section 5.7.

A part of the results presented in this chapter has also been published in [Baum12].

5.1 System Setup

5.1.1 Utterance Pre-processing

To implement the utterance pre-processing described in Section 3.3, several components were combined. Word transcripts were produced for the speech utterances with an automatic speech recognition system

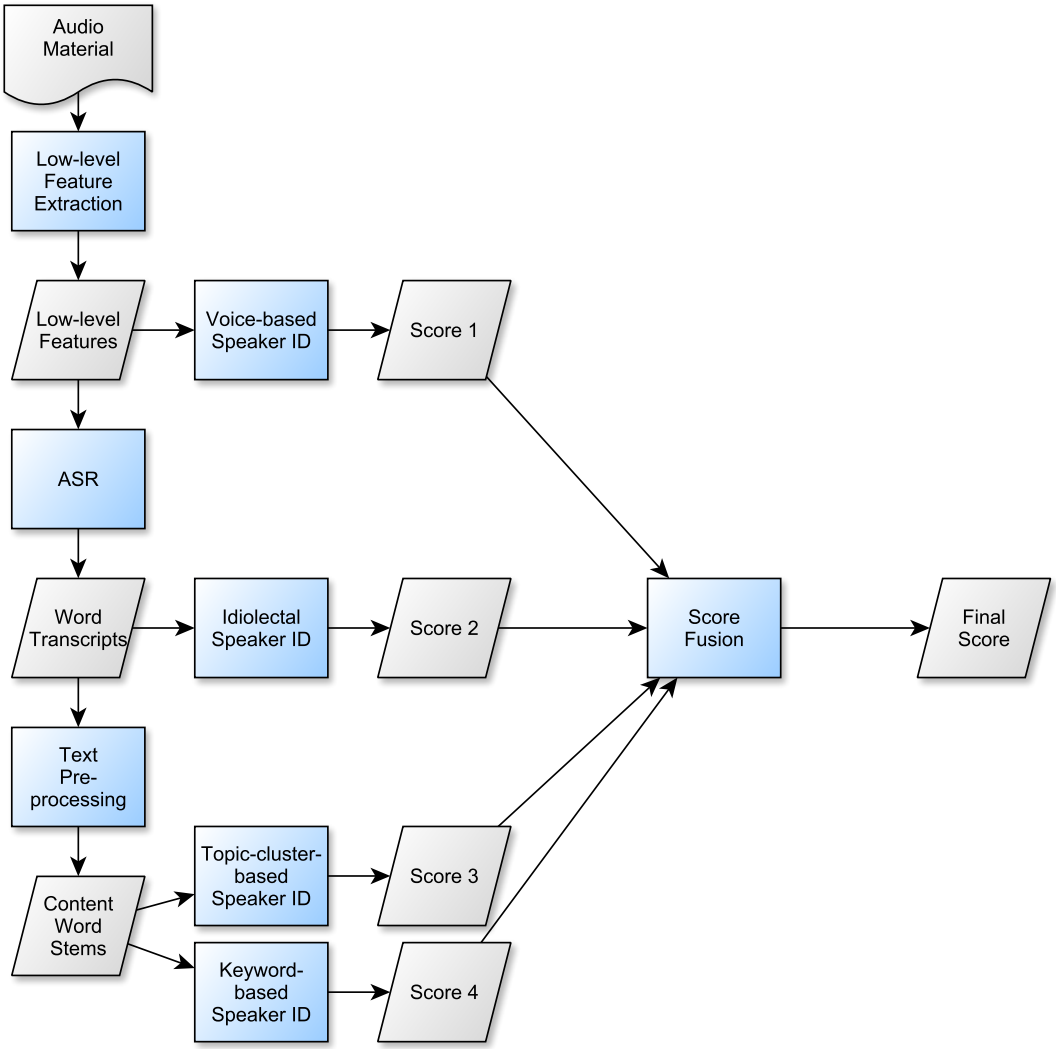


Figure 5.1: Processing chain for the final speaker recognition system that combines voice-, idiolect-, and topic-based recognition.

similar to the setup described in [Mertens09]. The system is based on the Julius[1] engine from ISTC [Lee01] and uses Fraunhofer IAIS models for German. The triphone acoustic models were trained on approximately 100 hours of German broadcast news data. The 3-gram language model was generated from a large corpus of German newswire data, amounting to approximately 150 million running words. The dictionary contains 200,000 words — which is necessary because of the prevalence of compounding in the German language. To give an estimate on the expected quality of the transcriptions: On a broadcast news corpus described in [Baum10a], the speech recogniser had a word error rate of 26% for planned speech with no background noise, of 33% for planned speech with talk in the background, and of 33% for spontaneous speech.

After ASR, stop words from a list of 204 very frequent German words were removed from the transcripts. Then, the remaining words were stemmed with the Snowball[2] stemmer for German [Porter01]. Stems that occurred less than twice in the combination of background set and speaker training set were discarded so that the number of stem types used was reduced to approximately 61,000.

5.1.2 Keyword-based System Setup

For keyword-based speaker recognition, Section 3.4 describes two scoring methods: log-likelihood ratio scoring and the Support Vector Machine classifier. The log-likelihood ratio scoring and the feature extraction of the SVM-based method were implemented in Perl. For log-likelihood ratio scoring, the adaptation coefficient α and discounting factor d which had been empirically determined for [Baum09a] were used: $\alpha = 0.98$ and $d = 0.45$. The speaker modelling and scoring of the SVM-based method were done using the SVMlight package[3] [Joachims99] to train a Support Vector Machine for each speaker and classify the test documents. The SVMs had a linear kernel and the default value was used for the regularisation parameter C (the inverse of the square of the average vector norm). As explained in Section 3.4.2, the misclassification of a positive (speaker) instance had to be penalised during training by factor $j = 3,690$, the ratio of background documents to speaker training utterances ($\frac{18,452}{5} \sim 3,690$).

5.1.3 Topic-model-based System Setup

The topic-model-based speaker recognition system made use of the Mallet package [4] [McCallum02] for finding the domain-inherent topics from the background set with Latent Dirichlet Allocation and for estimating the topic mixtures of the speaker utterances. As the number of topics T must be set in advance, the optimal T for the purpose of speaker recognition was determined in experiments described

[1] `http://julius.sourceforge.jp/`, last accessed: 2013-02-25
[2] `http://snowball.tartarus.org/`, last accessed: 2013-02-25
[3] `http://svmlight.joachims.org/`, last accessed: 2013-02-25
[4] `http://mallet.cs.umass.edu/`, last accessed: 2013-02-25

in Section 5.5.4. For the Bundestag corpus, $T = 20$ was found to give the best performance and is used throughout the experiments, if not stated otherwise. The topics were trained with 5,000 iterations of Gibbs sampling, the topic mixtures for the speaker utterances were estimated with 100 iterations. The Dirichlet prior parameters α and β were the default values of 50 and 0.01, respectively.

Like for the keyword-based system, SVMlight was used to train the speaker SVMs, this time with the Kullback-Leibler kernel from Equation 3.14, which was implemented as a user-defined kernel. Appropriate values for the scale and shift parameters a and b had to be found for the task of topic-based speaker recognition. Thus, a grid-search on the parameters space was done, as is recommended in [Hsu03], looking for good speaker recognition performance – as measured by equal error rate – on the development set. The parameter space was delineated by exponentially growing sequences of a and b, with $a \in \{2^{-12}, 2^{-10}, ..., 2^{-2}, 2^0\}$ and $b \in \{2^{-10}, 2^{-8}, ..., 2^2, 2^4\}$. As training models for all 235 speakers for each point in parameter space would have been too computationally costly, a subset of ten speakers was randomly selected. Only for these speakers, models were trained with all combinations of a and b, and only their development data was used for estimating the equal error rate of the resulting systems. The parameter set that gave the best equal error rate at the fastest processing time was picked, which was $a = 1$ and $b = 1$. The regularisation parameter was set to $C = 1$.

Also, SVMs with a Radial Basis Function kernel (see Equation 3.14) were trained as a baseline to compare the Kullback-Leibler kernel against, to see if it really is better than the default RBF recommended by [Hsu03]. The RBF parameters, $\gamma = 8$ and $C = 0.5$, were analogously determined, searching the parameter space delineated by $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^1, 2^3\}$ and $C \in \{2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}\}$ for the setup that produced the best EER with the least processing time.

As with the keyword-based system, the speaker utterances were used as positive examples and the background text documents as negative examples. The misclassification of a speaker example was also penalised by factor $j = 3,690$.

5.2 Baseline Systems

To estimate the performance of existing speaker recognition approaches on the new Bundestag corpus, two baseline systems were implemented: a spectral GMM-UBM system after [Reynolds00, Reynolds95a], and an idiolectal system with word-n-gram SVMs after [Shriberg05]. The GMM-UBM approach was chosen because it is widely used in the literature as the baseline reference, and thus makes the results in this work at least roughly comparable with the published figures. The idiolectal system was selected because using linear SVMs with word-n-grams is methodically similar to topic-based speaker recognition with SVMs and content word stems. It allows to assess the difference between the idiolectal features

based on function word n-grams and the topical features based on content word stems.

5.2.1 Spectral Speaker Recognition

The spectral system was a GMM-UBM system implemented in C++ based on [Reynolds00, Reynolds95a]. It made use of and added to the Fraunhofer IAIS iFinder SDK, which bundles implementations of various audio analysis methods. The system used Mel frequency cepstral coefficients (MFCCs) as features: 12 cepstral coefficients and first and second order derivatives were computed and cepstral mean subtraction was applied. The MFCC computation was done using the feature extraction implementation of the ISIP speech recognition software[5]. Voice activity detection was done by identifying homogeneous segments in the MFCC feature stream with Bayesian Information Criterion (BIC) segmentation [Tritschler99, Cheng10] (implemented in the iFinder SDK) and then classifying the segments into speech and non-speech. The classification was achieved by comparing the output of two GMMs, one for speech and one for non-speech, and assigning the class of the GMM that yields the higher probability for the segment's features. Only MFCCs labelled as speech were used by the speaker recognition system.

GMMs for the speakers were derived from a Universal Background Model (UBM) via *maximum a posteriori* adaptation (MAP). The UBM was combined from two gender-dependent GMMs (one trained on female and one on male speech) to achieve gender balancing. The gender-dependent GMMs had 512 mixtures each and were trained on 290 (female) and 545 (male) minutes of broadcast speech from an unrelated unpublished broadcast news corpus. They were then combined into one 1,024-mixture gender-independent UBM. The MAP adaptation was implemented in C++ according to [Reynolds00]. During recognition, processing was sped up by scoring each frame on the background model, finding the 10 best matching Gaussian components, and using only the 10 corresponding components of the speaker models for score computation. The speaker models were MAP-adapted with 2 minutes of audio data from each of the 5 speaker training files, i.e., with 10 minutes of speaker training data.

5.2.2 Idiolectal Speaker Recognition

The idiolectal speaker recognition baseline uses trigrams of frequent words as features and linear kernel SVMs as classifiers, like in [Shriberg05]. Like for the topic-based systems, SVMlight was used to train the speaker SVMs and do the scoring. In contrast to the topic-based system, no utterance pre-processing, no stemming and stop word removal, was applied. This is because stop words are important function words and function word trigrams are the features that are supposed to capture the style and idiolect of the speaker. Removing the stop words greatly reduced system performance for the idiolectal system

[5] `http://www.isip.piconepress.com/projects/speech/software/downloads/`, last accessed: 2013-02-25

in a preliminary experiment, almost doubling the EER (from 18% to 33%). Accordingly, all words in the transcripts and texts were retained and combined into word trigrams. Trigrams occurring less than twice in the speaker training and background material were discarded for efficiency reasons, thus the final trigram dictionary contained approximately 900,000 different trigram types. After that, tf-idf vectors were calculated for each utterance, where the terms were the trigrams. Tf-idf term weighting was used instead of the rank normalisation scheme from [Shriberg05] because it gave better results also for idiolectal speaker recognition in preliminary experiments. Then, linear kernel SVMs were trained with the speaker utterances as positive examples and the background text documents as negative examples, again penalising misclassification of a speaker example more strongly by factor $j = 3,690$, the ratio of background documents to speaker training utterances ($\frac{18,452}{5} \sim 3,690$). Note that the lack of genuine spoken utterances as background training material may decrease the idiolectal system's performance, but that the same background set had to be used.

5.3 System Fusion

The topic-based systems were fused with the spectral and the idiolectal system in order to see whether they provide useful additional information for speaker discrimination. The fusion was done as a linear combination of the scores. The optimal system weighting was determined on the development set by Linear Logistic Regression with the FoCal package[6] [Brümmer06].

The usefulness of test utterance length as side information in the fusion was also tested. For this, a variant of the FoCal toolkit inspired by [Ferrer08] called FoCal Bilinear[7] was used, which fuses system scores taking side information into account. Side information should describe the test utterances as belonging to one or more categories which can have an influence on the fused systems' reliability. In the experiments described in Section 5.6.4, the categories are "short utterances" and "long utterances" because topic-based speaker recognition is more reliable for long utterances (see Section 5.6.3). The FoCal Bilinear toolkit expects side information in the form of (soft) category membership labels for each trial, i.e. one probability distribution per trial, indicating the likelihoods of belonging to the utterance categories. It trains bilinear fusion weights with Linear Logistic Regression – bilinear meaning that if the side information is constant, the fusion is a linear combination of scores and vice versa. During fusion, subsystem scores are weighted by the trained fusion coefficients and multiplied with the side information to improve the fused system's reliability.

[6] `https://sites.google.com/site/nikobrummer/focalbinary`, last accessed: 2013-02-25
[7] `https://sites.google.com/site/nikobrummer/focalbilinear`, last accessed: 2013-02-25

5.4 Significance Testing for Evaluation Results

To verify that the differences in equal error rates which are presented in this chapter are indeed statistically significant, the systems' performance was compared in pairwise McNemar's tests [Gillick89, McNemar47].

McNemar's test compares the performance of two algorithms and indicates, whether a difference between them is statistically significant. It works by comparing the numbers of trials where one system made a false decision while the other one was correct. That is, for systems A and B, it looks at the number of trials that A classified correctly and B misclassified, and vice versa. The null hypothesis is that the probability of system A making an error in a specific trial while system B does not is equal to the probability of system B making an error while system A does not. To facilitate the pairwise test, a 2×2 contingency table is drawn up, showing how many trials were correctly classified by both, incorrectly classified by both, correctly classified by A but incorrectly classified by B, and correctly classified by B and incorrectly classified by A – see Table 5.1 for an example. (Strictly speaking, the cells in the table where both systems agree are not used in the test.)

		System B	
		Correct	Incorrect
System A	Correct	60,627	8,585
	Incorrect	19,118	8,490

Table 5.1: Example 2×2 contingency table, showing the number of correct/correct, correct/incorrect, incorrect/correct, and incorrect/incorrect trials for systems A and B.

Given the data from the table, the null hypothesis that both systems are equally likely to make an error for a given trial can be examined with a two-tailed test comparing the hypothetical and the real distribution of errors. This is done by computing a test statistic and comparing it with the χ^2 distribution (for a detailed explanation see [Gillick89, McNemar47]). This yields a p-value which is compared with the significance level α – the null hypothesis is rejected if the p-value is less than α.

As the main performance measure used in this thesis is the equal error rate, the differences in EERs were significance tested. To obtain the necessary 2×2 tables, the scores output by the systems had to be rendered into hard decisions using decision thresholds. The systems' decision thresholds were tuned on the development set to the operating point which gave equal error performance. Then, the numbers of trials where one system made a false decision while the other one was correct were compared with McNemar's test. The significance level used was $\alpha = 0.01$, thus the null hypothesis was rejected if the obtained p-vale was less than 1%.

5.5 Calibration

Before the final evaluation of the proposed methods (described in the next section), the systems were calibrated to select the best-performing approaches from the proposed ones and to optimise the parameters. Also, a number of assumptions made in Chapter 3 were verified. In the experiments described in this section, models were trained with the Bundestag training data, while the evaluation was done on the development part of the corpus. In the final evaluation, the trained models were evaluated on the test data set. This separation into development and test was done to avoid over-fitting the system setup to the test material. Each development utterance was scored against all speaker models, giving one true speaker trial and 234 impostor trials per utterance.

5.5.1 Modelling for Keyword-based Speaker Recognition

Section 3.4 describes two different training and scoring methods for keyword-based speaker recognition: Log-likelihood Ratio scoring (Section 3.4.1) and Support Vector Machine Classifiers (Section 3.4.2). In order to compare their performance, both methods were implemented and compared on the development dataset. Both used the utterance pre-processing described in Section 3.3, as stemming and stopword-removal had proved beneficial for the performance of log-likelihood scoring in the experiments for [Baum09a]. The parameters for both systems are detailed in Section 5.1.2. Additionally, the SVM-based system used tf-idf term weighting.

Figure 5.2 shows their performances in a DET-plot, the equal error rates are 17.7% for the SVM-based system and 18.0% for the Log-likelihood Ratio scoring method. The difference in EER is statistically significant at $\alpha = 0.01$ level. As the SVM-based method has a slightly lower EER, and as it performs better at the high-miss-probability-end of the DET curve, it was chosen as the preferred training and scoring method for keyword-based speaker recognition. It was thus used in all the following experiments.

5.5.2 Influence of Pre-processing on Keyword-based Speaker Recognition

It is assumed in Section 3.3 (and already transpired in experiments in [Baum09a]) that utterance pre-processing reduces the difference between written text and ASR transcripts, compacts the topic information in an utterance, and thus serves to improve topic-based speaker recognition.

This was verified in an experiment: Two SVM-based keyword-based speaker recognition systems were set up, one with the utterance pre-processing described in Section 3.3, the other using the full utterances, which were only normalised by removing punctuation and non-alphabetic characters. In both systems, only words that occur twice or more in the speaker training and background sets are taken into account, for system efficiency reasons. Also, both used tf-idf term weighting.

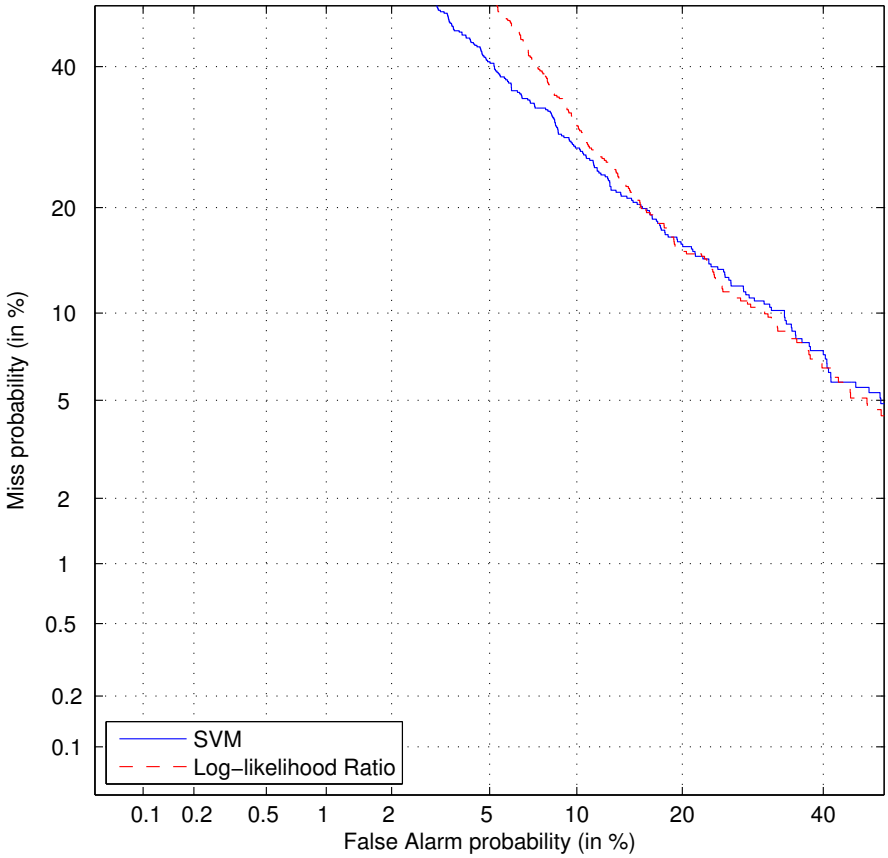


Figure 5.2: DET-curves for keyword-based speaker recognition with Log-likelihood Ratio scoring and Support Vector Machine Classifiers.

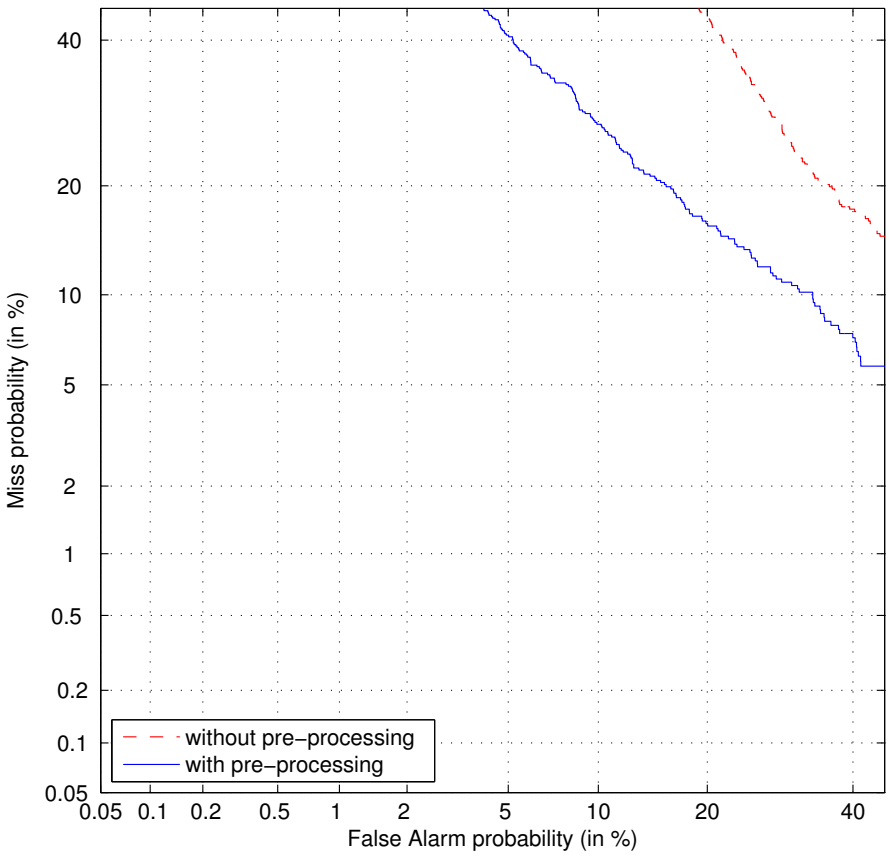


Figure 5.3: DET-curves for keyword-based speaker recognition with and without utterance pre-processing.

The performance of the resulting systems differed quite substantially, as can be seen in the DET curves in Figure 5.3. The equal error rate is improved from 28.5% to 17.7% by utterance pre-processing, a statistically significant difference (at $\alpha = 0.01$). This backs up the assumption that utterance pre-processing benefits the system performance by compacting the topic information used by the proposed system and by removing differences between written text and ASR transcripts. Thus, for all other topic-based systems mentioned in this chapter, text pre-processing was used.

5.5.3 Influence of Term Weighting on Keyword-based Speaker Recognition

Section 3.4.2 describes two possible weighting methods for the content word stem frequencies which are used as features in keyword-based speaker recognition with SVM classifiers: tf-idf and rank normalisation. The weighting is necessary to relate the frequency of a stem in the speaker training material to its frequency in the background set and thus to determine whether the stem is characteristic for the speaker. A stem that occurs frequently in a speaker's utterances but rarely in the background material should carry more information about the speaker than a stem that occurs frequently in both — and should be

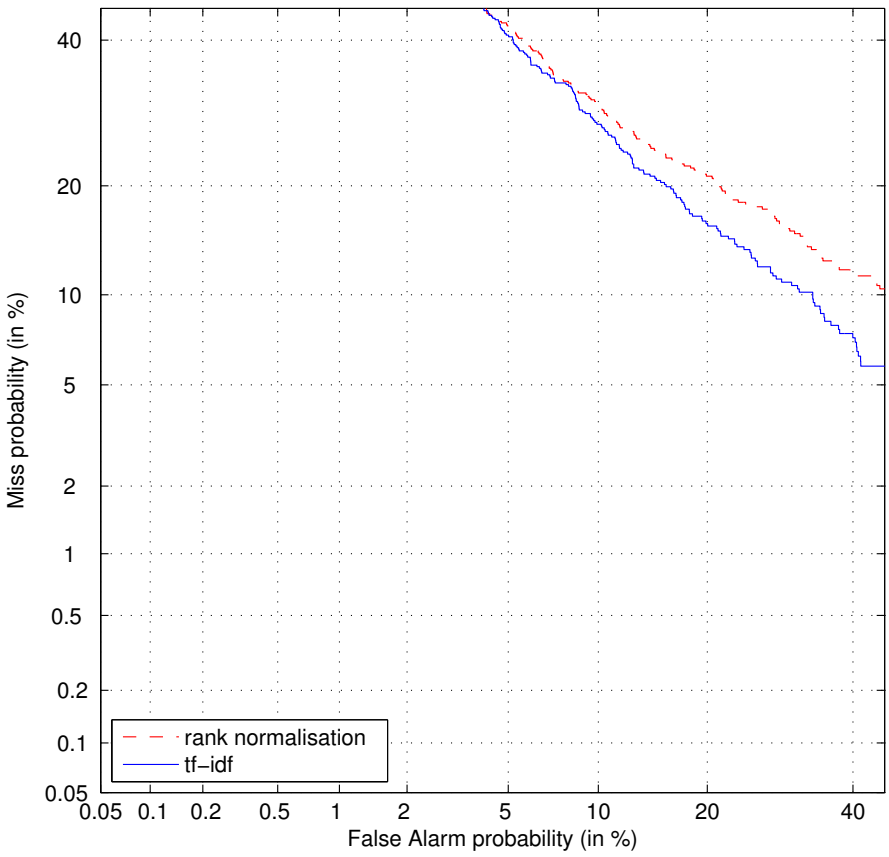


Figure 5.4: DET-curves for keyword-based speaker recognition with rank normalisation and tf-idf term weighting.

given more emphasis accordingly.

Rank normalisation is used in the idiolectal speaker recognition system in [Shriberg05], while in text categorisation and authorship attribution, tf-idf seems to be more popular [Joachims98, Diederich03].

In order to see which weighting scheme is more appropriate for keyword-based speaker recognition, two SVM-based systems were set up which differed only in term weighting. Both used pre-processing as described in Section 3.3 and the parameters detailed in Section 5.1.2. Figure 5.4 shows their performances in a DET-curve. Using tf-idf instead of rank normalisation improved the equal error rate from 20.7% to 17.7%, again a statistically significant difference (at $\alpha = 0.01$). Thus tf-idf was employed for the keyword-based systems in all other experiments.

5.5.4 Optimal Number of Topics for Topic-model-based Speaker Recognition

As mentioned in Section 3.5, the Latent Dirichlet Allocation algorithm used to find the domain-inherent topics from the background corpus needs as input the number of desired topics T. For the Bundestag corpus, T was experimentally determined by setting up a number of topic-model-based speaker recogni-

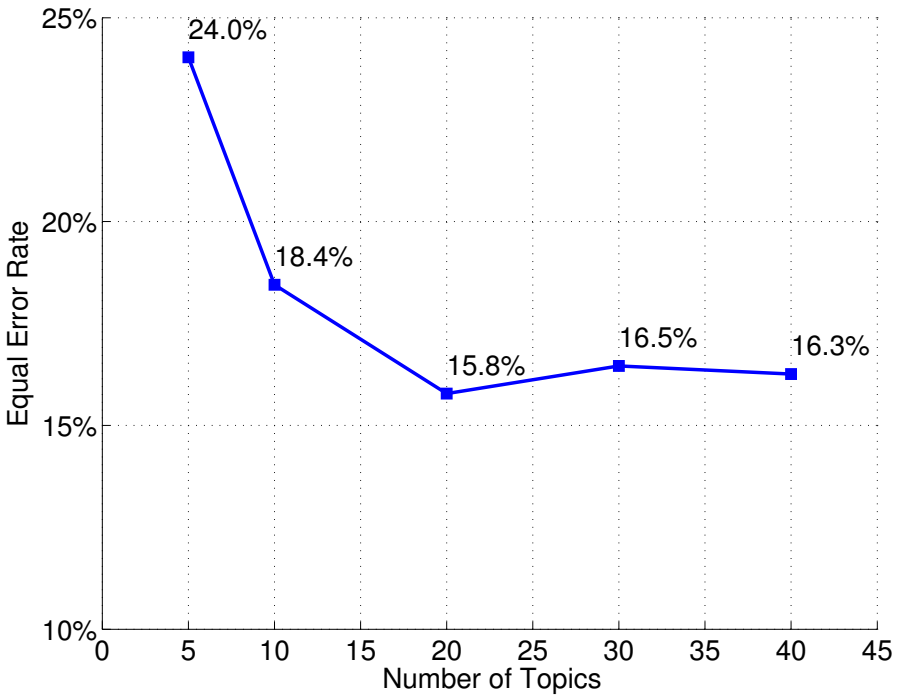


Figure 5.5: Equal error rates for topic-model-based speaker recognition for different numbers of topics.

tion systems with speaker prototype models, differing only in the number of topics. The performance of these systems was evaluated and compared. Figure 5.5 shows that the equal error rate decreases with increasing T until T reaches 20, at which point it stagnates. The differences in EER are all statistically significant (at $\alpha = 0.01$), except between 30 and 40 topics. Thus, 20 topics were used for the topic-model-based systems in all other experiments. When applying topic-model-based speaker recognition to other domains, this step presumably has to be repeated for the new background training data.

5.5.5 Distance Measure for Topic-model-based Speaker Recognition

To measure the similarity between an utterance's topic and the speakers' preferences, a measure of distance between topic mixtures is required, see Section 3.5.4. Four well-known distance measures were compared to select the one best suited for topic-model-based speaker recognition: Cosine, Bhattacharyya, Jensen-Shannon, and Kullback-Leibler divergence / distance. Again, topic-model-based speaker recognition systems with speaker prototype vectors were created, which differed only in the distance measure used. The results are given in Figure 5.6 and Table 5.2.

The overlapping DET curves in Figure 5.6 indicate that the performance of the distance measures is similar, yet the difference is statistically significant in all pairings (at $\alpha = 0.01$) according to McNemar's test. Kullback-Leibler divergence has a the best EER on the development set, thus it was selected as the scoring method for the other experiments.

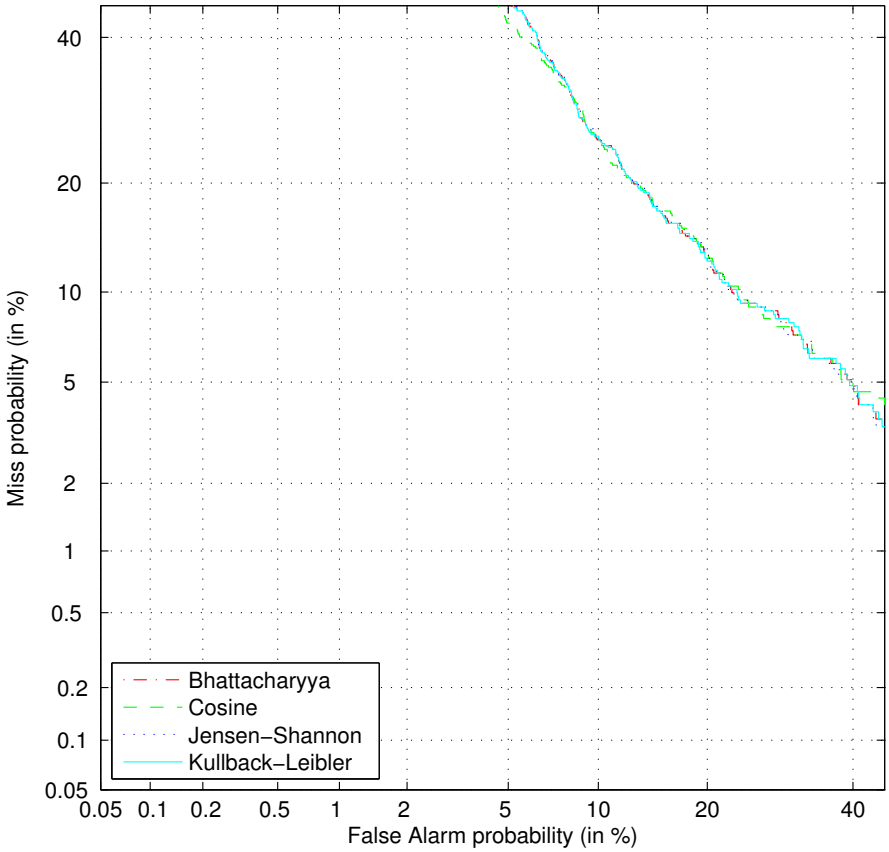


Figure 5.6: DET-curves for topic-model-based speaker recognition for different distance measures used in scoring.

Distance measure	EER
Cosine distance	16.3%
Jensen-Shannon divergence	16.1%
Bhattacharyya distance	16.0%
Kullback-Leibler divergence	*15.8%*

Table 5.2: Equal error rates for different distance measures.

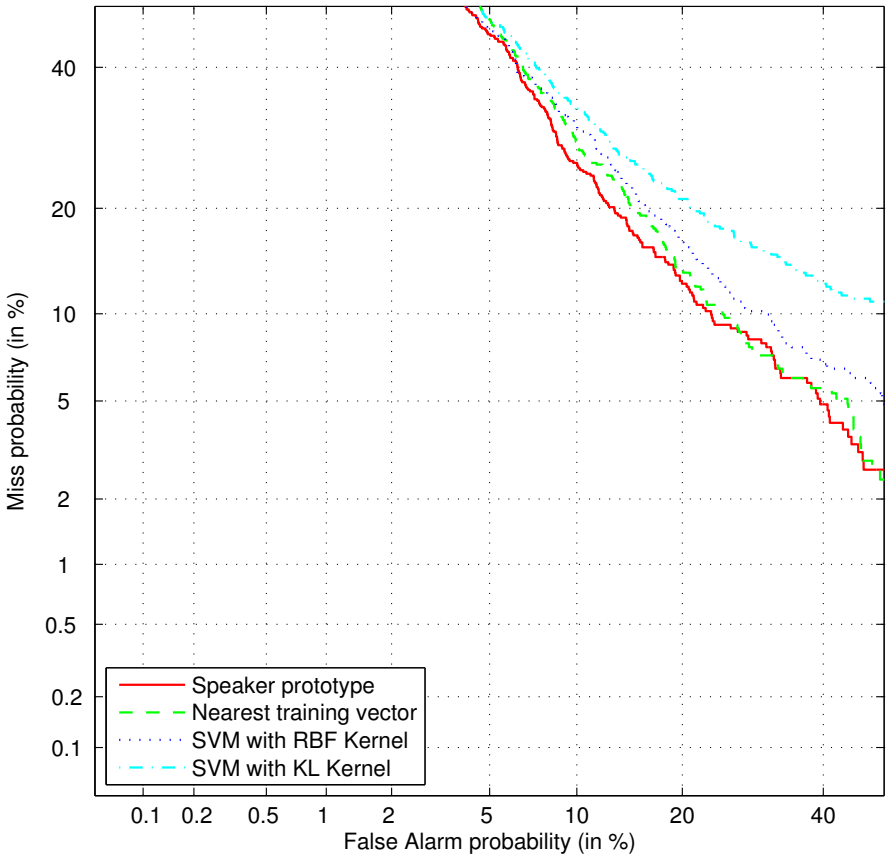


Figure 5.7: DET-curves for topic-model-based speaker recognition with different classifiers.

5.5.6 Modelling for Topic-model-based Speaker Recognition

Three possible speaker modelling and scoring methods are proposed in Sections 3.5.4 through 3.5.6: using speaker vector prototypes, lazy learners, and Kullback-Leibler kernel Support Vector Machines. To see which one is best suited for topic-model-based speaker recognition a further experiment was conducted in which only the modelling and scoring was changed. Four systems are compared: the proposed three methods and another SVM system with a Radial Basis Function kernel, to see whether the specialised KL kernel gives better performance than the standard RBF kernel. Figure 5.7 and Table 5.3 show the results. The differences in EER are statistically significant (at $\alpha = 0.01$).

It appears that the simple speaker vector prototype modelling, which averages the speakers' topic

Classifier	**EER**
SVM with KL Kernel	20.7%
SVM with RBF Kernel	18.2%
Nearest training vector	17.2%
Speaker prototype	*15.8%*

Table 5.3: Equal error rates for topic-model-based speaker recognition with different classifiers.

preferences from their training utterances, gives the best performance. The lazy learner, which for scoring uses the speaker's training vector that is closest to the test utterance, performs worse. This indicates the assumption that a speaker is a likely source for an utterance if he or she was observed once to talk about the same topic does not give good results in practise, at least not for the Bundestag data. It is preferable to instead use average speaker interests.

Interestingly, the SVM systems perform worse than both simpler modelling methods. It is likely that the very small number of positive training examples per speaker is just too little data to train really good discriminative speaker classifiers based on the topic-mixture feature vectors. Their performance would probably improve with more training examples for the speakers, however, those may not be available in the chosen domain. Also, the standard RBF kernel outperforms the specialised KL kernel. This may be the case because of the additional kernel parameter in the KL kernel: While for the RBF kernel only γ and C are to be experimentally determined, the KL kernel has three parameters, a, b, and C (see Section 5.1.3 for a description of the parameter estimation). As the parameter optimisation had to be done by observing the speaker recognition error rates estimated with only 10 test speakers, it is possible that both systems would give better results with different parameters, estimated on a larger development set. However, it is harder to find the best parameter combination for the KL kernel, because the parameter space has an additional dimension.

As the speaker prototype vector approach gives the best results on the development data, this modelling and scoring method was used in the other experiments.

5.5.7 Final System Configuration for the Topic-based Systems

The calibration of individual system parameters on the development set yielded the optimal settings described in the previous sections. As an exhaustive search over the "parameter space" of all possible system configuration combinations would be too computationally costly, the combination of the individually best parameter settings is used as the final system configuration in the evaluation described in the next section.

For the topic-model-based system it was found that using 20 topics (compared to 5, 10, 30, and 40) gave the best results, that speaker classification should be done with prototype vector modelling (in-

stead of lazy learning or SVMs), and that Kullback-Leibler divergence was the best-performing distance measure (compared with Cosine distance, Jensen-Shannon divergence, and Bhattacharyya distance).

For the keyword-based system, it was verified that pre-processing (with stop word removal and stemming) indeed strongly increases performance, that speaker classification should be done with SVMs (instead of Log-likelihood Ratio scoring) and that for term weighting tf-idf works better than rank normalisation.

5.6 Evaluation Results

After verifying the basic assumptions on which topic-based speaker recognition is based and optimising the set-up of the implementation, the main evaluation was conducted to gain some insight on the research questions of this thesis: Is it possible to recognise speakers from topic information? How do the different representations of speaker topic preferences influence performance for topic-based speaker recognition? How does topic-based speaker recognition compare with methods based on other cues? Does performance improve by combining several cues? How much does topic-based speaker recognition depend on circumstances such as the length of the test utterance or the age of the speaker training material?

The experiments were conducted on the test set of the Bundestag corpus to see how well systems tuned on the development data perform on a different data set. Each test utterance was compared against all speaker models, giving one true speaker trial and 234 impostor trials per utterance.

First, the topic model obtained by Latent Dirichlet Allocation was subjectively evaluated to see whether the automatically derived topics make sense to humans and whether politicians with a known topic preference are assigned to the corresponding topic cluster (Section 5.6.1). Then, the performance of the keyword-based, topic-model-based, idiolectal, and spectral systems were evaluated on the test data (Section 5.6.2). The influence of the length of the utterance on the two topic-based speaker recognition methods is investigated in Section 5.6.3. Finally, topic, idiolect, and voice information were fused to see if this combination of different cues leads to improved speaker recognition performance (Section 5.6.4).

5.6.1 Subjective Evaluation of Speaker-Topic Assignment

After training the 20 topic model for topic-model-based speaker recognition with Latent Dirichlet Allocation on the Bundestag domain, it was subjectively evaluated. This was done to see whether the major domain topics found by unsupervised "clustering" in the unlabelled topic training data are coherent and make sense to a human. After topic model training, the resulting topics (clusters) were manually labelled by a human subject according to their most probable words and most representative documents. It was found that of the 20 topics 5 are concerned with organisational issues (announcements, votes, short ques-

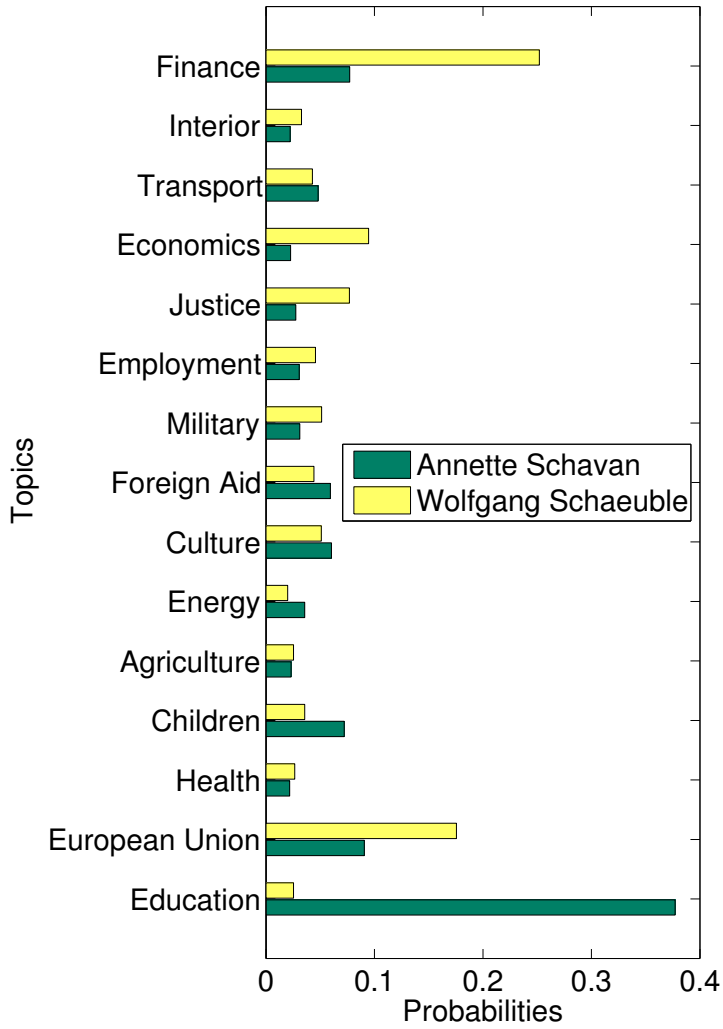


Figure 5.8: Topic mixtures for the ministers of finance (Wolfgang Schäuble) and education and research (Annette Schavan), with manually labelled topics.

tions, etc.). The other 15 topics are "content topics", corresponding to different political departments: education, Europe (EU), health, children, agriculture, energy, culture, foreign aid, military, employment, justice, economics, transport, interior, and finance.

After the manual topic labelling, it was checked whether politicians with known topic preferences, such as ministers, were assigned the expected automatically derived topics (for the time period of 2009/2010). For example, Figure 5.8 shows the speaker prototype vectors for the content topics of the ministers of finance (Wolfgang Schäuble) and education and research (Annette Schavan). As expected, the ministers have strong assignments to the topics of their department. In a short qualitative analysis, this seemed to be true for many speakers. This supports the hypothesis that the speakers' topic preferences can be determined in an automatic, unsupervised fashion. A quantitative evaluation of the human appraisal of the automatically derived topics, like in [Chang09], and of the speaker/topic assignments would be

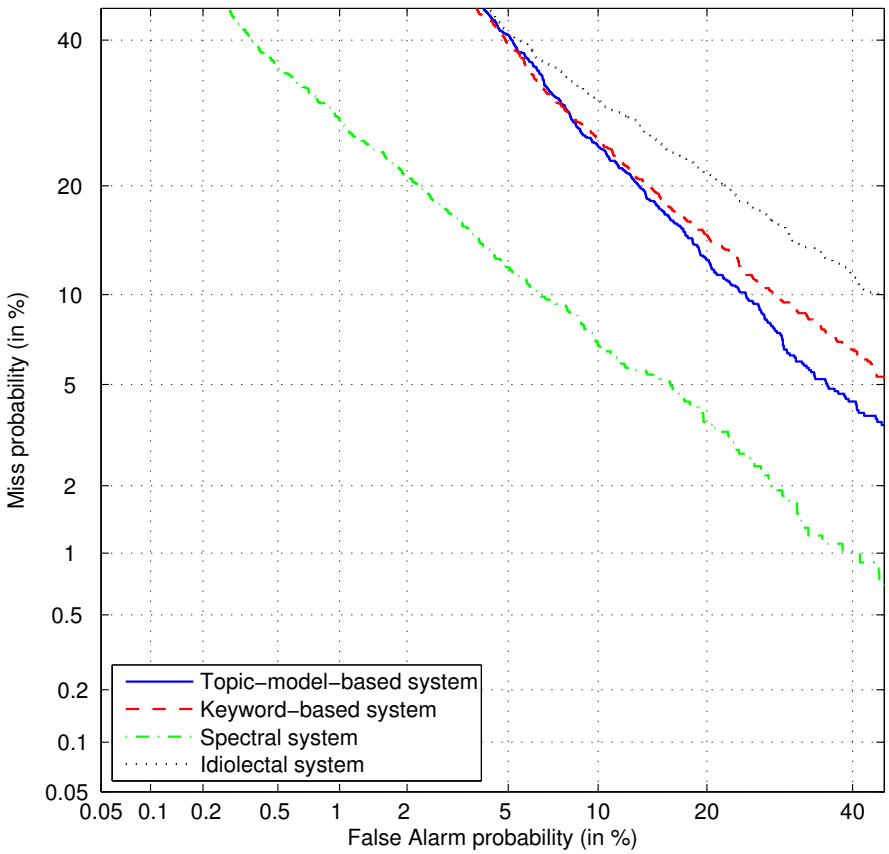


Figure 5.9: DET-curves comparing the spectral, idiolectal, keyword-based and topic-model-based speaker recognition systems.

interesting as future work.

5.6.2 Performance Comparison of the Different Speaker Cues

To answer the question if topics can be used for speaker identification and to estimate how this cue compares with other cues, the optimised versions of the two topic-based systems (described in Chapter 3) and the two baseline systems (described in Section 5.2) were evaluated on the Bundestag test set. The DET-curves are shown in Figure 5.9, the equal error rates (EERs) are listed in Table 5.4. Note that the differences in EERs are statistically significant (at $\alpha = 0.01$), apart from the comparison between the keyword-based and the topic-model-based system. This means that the spectral system performs significantly better than the topic-model- and keyword-based systems, which in turn perform significantly better than the idiolectal system.

The results indicate that topic information can indeed be used for speaker recognition, although it is not as reliable a cue as the speaker's voice captured by the spectral features. This is intuitive, as several speakers may well be experts for the same topics, thus the discriminative power of the features is not as

System	EER
Idiolectal	20.8%
Keyword-based	17.0%
Topic-model-based	16.3%
Spectral	8.6%

Table 5.4: Equal error rates for the spectral, idiolectal, keyword-based, and topic-model-based speaker recognition systems.

high as with voice recognition.

For the Bundestag corpus, topic-based speaker recognition outperforms the idiolectal approach. This probably stems partly from the fact that the ASR transcripts are not error-free and the background training material is written text. The idiolectal system should be particularly affected by this difference, as it uses the information contained in the frequencies of short function words, which are frequently misrecognised by ASR. Because the topical system uses the longer content words, its performance is not degraded as much by ASR errors.

It is also noteworthy that the topic-model-based system performs as well as the keyword-based system, although its features have much lower dimensionality: 20-dimensional topic mixtures compared to approximately 61,000-dimensional sparse tf-idf vectors (where the speaker training documents have 277 terms with non-zero weight on average). Apparently, it really are the topics that carry important information about the speakers and not just speaker-specific marker words like names and titles.

Finally, the contingency table comparing the spectral and the topic-model-based system (see Table 5.5) gives a first indication on whether the spectral approach could benefit from a fusion with a topic-based system. The table was used in McNemar's significance test on the performance difference between the spectral and the topic-model-based system (compare Section 5.4). While it shows that the spectral system clearly performs better, it also indicates that there are a high number of trials (17,401) where the spectral system could benefit from a fusion with the topic-model-based system.

		Topic-model-based	
		Correct	Incorrect
Spectral	Correct	178,688	33,312
	Incorrect	17,401	4,189

Table 5.5: Contingency table showing the number of correctly and incorrectly classified trials for the topic-model-based and spectral speaker recognition systems.

Manually inspecting a very small subsample of those 17,401 trials, it seemed that the spectral system confuses speakers with similar voice characteristics – as is to be expected. To get a quantitative estimate of this effect, speaker gender was used as an indicator of similar voice characteristics, based on the assumption that female voices sound more similar to other female voices than to male voices, and vice

versa. Of the 17,401 trials where the spectral system was wrong while the topic-model-based system was right, 91% (15,980) are same gender trials. For comparison: As the male/female ratio among speakers is 70 : 30, 58% of the total number of trials (233,590) are same gender trials. Of the 33,312 trials where the topic-model-based system was wrong and the spectral system was right, 54% (18,119) are same gender trials. Thus, the spectral system confuses same gender speakers much more often than at chance level, while the topic-model-based system does not. This reinforces the assumption that the spectral system mistakes similar sounding speakers for each other and the topic-model-based system does not. Thus, fusion can be expected to improve the overall systems' performance.

5.6.3 Influence of the Length of the Test Utterance on Topic-based Speaker Recognition

The performance of the topic-based speaker recognition systems can be expected to depend on the length of the test utterance. The assumption is that a longer utterance contains more topic-carrying words and allows for a better implicit or explicit estimation of the topic. This dependency can be evaluated on the given data by artificially decreasing the number of words in the test utterances. However, it is not immediately clear how the artificial utterances should be constructed from the original test material. Two possibilities were explored, mimicking two use cases: First, up to N_{max} words from the *beginning* of the speeches were used as test utterances, simulating a listener in parliament who forms a speaker hypothesis after hearing a few sentences of the speech. Second, up to N_{max} words were cut from the *middle* of the speech, simulating a quote used as a sound bite, for example in broadcast news. The artificially generated test utterances could at most be of length N_{max} and were shorter than that if the uncut utterances where shorter than N_{max}. Only content words (the words that remain after pre-processing) were taken into account here.

Table 5.6 and Figure 5.10 summarise the performance of the two topic-based speaker recognition systems, depending on the number of content words per test utterance and the position from where they were extracted. The average number of words per document, N_{avg}, is lower than the cut-off length N_{max} for large N_{max} because then most original documents are shorter than N_{max}.

From the results it appears that the performance indeed depends on the length of the test utterances and also (but not as strongly) on the extraction position. Satisfactory performance can be achieved when there are at least about 100 content words in the test utterance. If fewer words are available, the error rate increases rapidly. The topic-model-based approach seems to be much more sensitive to short utterances than the keyword-based system, while it performs better on long utterances. Interestingly, the difference between the two systems on short utterances depends on the extraction position: when

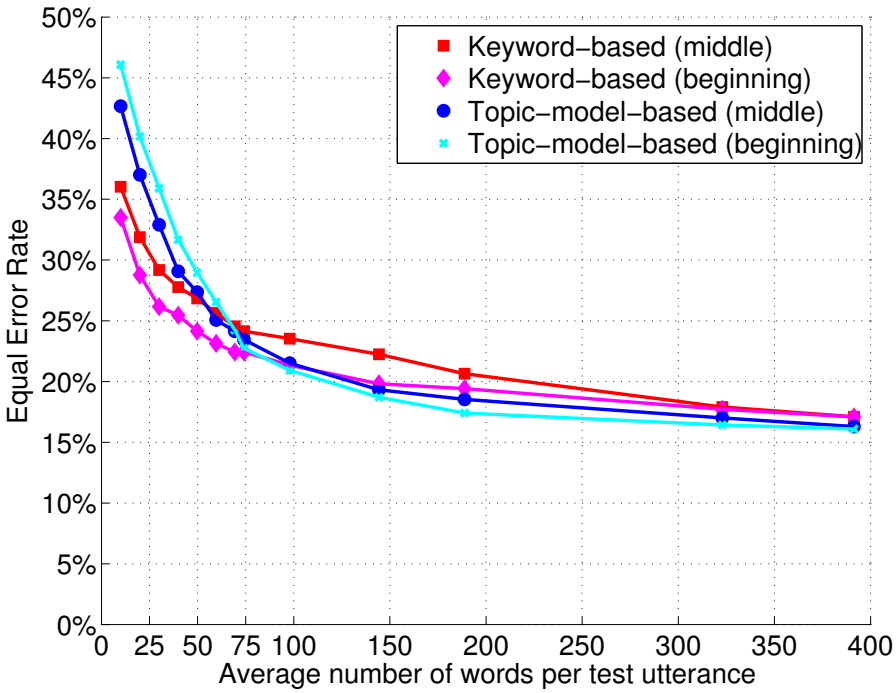


Figure 5.10: Influence of the average test utterance length and extraction position on the equal error rate.

N_{max}	N_{avg}	EER topic-model-based (beginning)	EER keyword-based (beginning)	EER topic-model-based (middle)	EER keyword-based (middle)
10	10.0	46.1%	33.5%	42.7%	36.0%
20	20.0	40.2%	28.8%	37.0%	31.9%
30	30.0	35.9%	26.2%	32.9%	29.2%
40	39.9	31.7%	25.5%	29.1%	27.8%
50	49.8	29.0%	24.1%	27.4%	26.8%
60	59.6	26.5%	23.1%	25.1%	25.7%
70	69.3	24.3%	22.4%	24.1%	24.6%
75	74.1	22.7%	22.4%	23.4%	24.1%
100	98.0	20.9%	21.3%	21.5%	23.5%
150	144.3	18.7%	19.8%	19.3%	22.2%
200	188.8	17.4%	19.4%	18.5%	20.6%
400	322.9	16.4%	17.7%	17.0%	17.9%
800	391.3	16.1%	17.1%	16.3%	17.1%
all	407.4	16.3%	17.0%	16.3%	17.0%

Table 5.6: Topic-based speaker recognition systems' equal error rates according to the number of words per test utterance and position where those words are extracted.

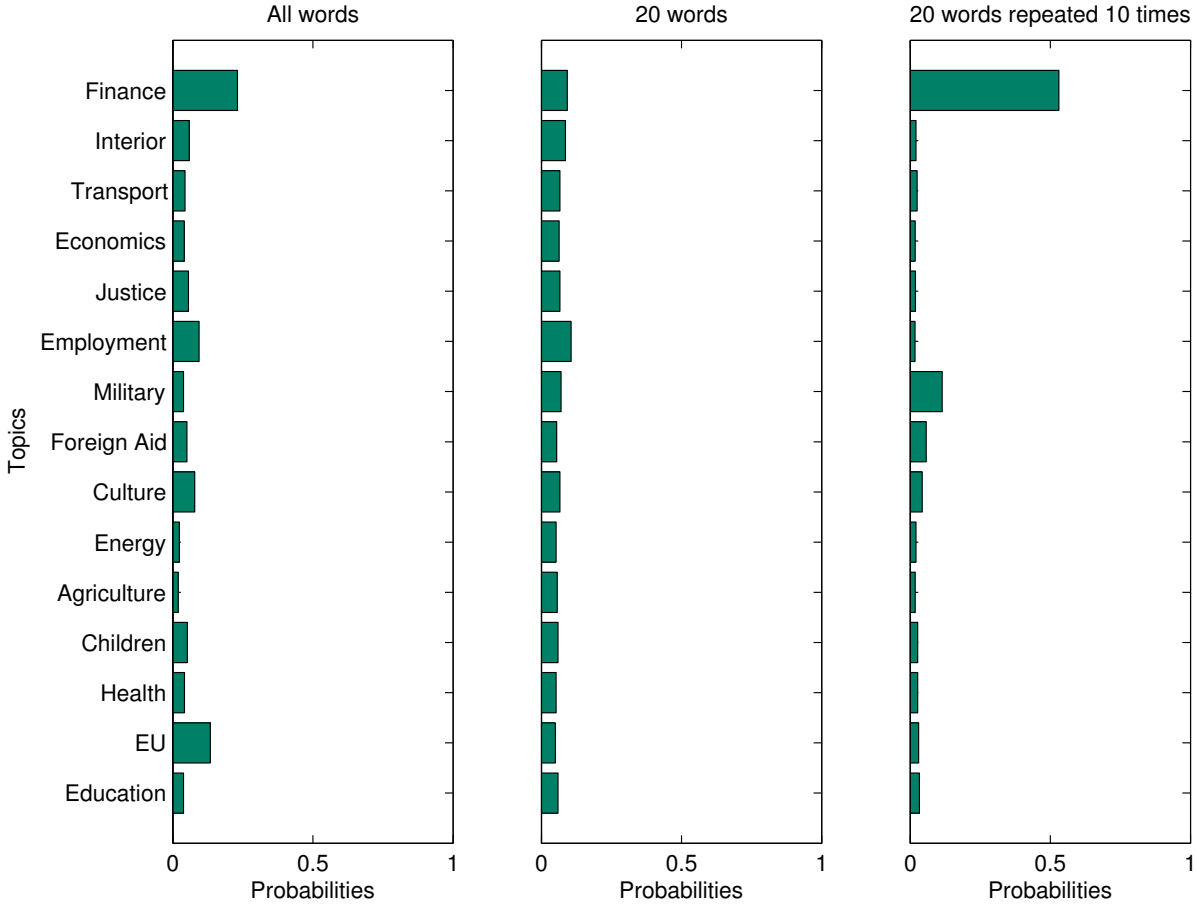


Figure 5.11: Topic mixtures for an example utterance using all 655 words (left), 20 words form the middle, and the middle 20 words repeated 10 times.

using the beginning of the speeches it is larger than if using excerpts from the middle.

This points to one of the reasons for the performance gap between both methods: because of its discriminative classification, the keyword-based system can put much more emphasis on speaker-specific marker words, and it only has to observe few of those. A special case of marker words that occur very often at the beginning of speeches are the speakers' names and positions, mentioned during the introduction. If the politicians' names are removed from the speeches, the EER of the keyword-based system on 20 words from the beginning increases from 28.8% to 30.6%, while the EER of the topic-model-based system only slightly changes from 40.2% to 40.7%. Obviously, the speakers' names are a very good hint for determining their identity — however, this is not an identification based on topic.

A second reason for the performance difference is that the topic mixture estimation requires a larger number of observed words. This is because the Dirichlet prior on the topic mixtures smoothes the estimated mixtures. In the described experiments, Dirichlet hyperparameter $\alpha = 50$ was used, corresponding to a prior observation count of 50 for each topic in each document. So, naturally, the few observations of

each topic in the short utterances do not change the prior sufficiently, and the estimated mixture is very similar to the prior. There are two ways to increase the impact of each observed word: The Dirichlet parameter α can be lowered to reduce the smoothing on the topic mixtures, but this would require re-training the topic model. An alternative is to artificially increase the observation counts of the words in the test utterances by repeating the utterances multiple times. The latter method was tested because it did not require re-training the topic model and thus the results are directly comparable to the previous ones. Figure 5.11 illustrates the principle: the leftmost topic mixture represents a full-length example utterance (with 655 words). The mixture in the middle was estimated on 20 words from the middle of the utterance, yielding a distribution similar to the uniform prior. The rightmost mixture was estimated on the middle 20 words repeated 10 times, artificially creating a 200 word utterance. Artificially extending short utterances can increase performance for topic-model-based speaker recognition. For example, the system's EER decreased from 40.2% to 34.1% on the 20 words (from the beginning) test utterances when they were repeated 10 times, artificially creating 200 word utterances. Thus, the repetition of short utterances (or a lower α) can be used to increase performance for topic-model-based speaker recognition on corpora with short utterances. In the Bundestag corpus, however, the vast majority of utterances (94%) have more than 100 content words. Yet, the topic-model-based system with artificially extended utterances still performs worse than the keyword-based system's EER of 28.8% for 20 word utterances and 30.8% for 20 word utterances without names.

The rest of the difference may be due to other speaker-specific marker words, and it remains a question for future research what the precise attributes of these words are — and whether they are subject to change over time or highly characteristic for the speaker.

To sum up: both systems' performance depends on the utterance length, with 100 words yielding satisfactory results. The keyword-based system can make use of speakers' names mentioned in the utterances and performs better for short utterances. For longer utterances, the topic-model-based system gives better results. As the length of the utterance seems to be a rough measure of the systems' reliability, using it as side information for system fusion might improve fused system performance. This hypothesis was tested in the next section.

5.6.4 Fusion of Topic-based Systems with Baseline Systems

To answer the question whether topic features provide useful additional information in a speaker recognition system when combined with other cues, the two topic-based and the two baseline systems were fused and the results compared. The fusion was done by a linear combination of the systems' scores (see Section 5.3). The systems' weights for this combination were estimated with Linear Logistic Regression

on the development data. Using these weights, the individual systems' scores on the test data could be fused to obtain new score values. Various combinations were tried to see how much information each system adds, Table 5.7 shows the obtained equal error rates and whether the differences are significant at $\alpha = 0.01$ in pairwise McNemar's tests.

ID	**System**	**EER**	**significantly better than**
a	Idiolectal	20.8%	
b	Keyword-based	17.0%	a
c	Topic-model-based	16.3%	a
d	Spectral	8.6%	c,b,a,e,f,g
e	Idiolectal + keyword	14.9%	c,b,a
f	Idiolectal + topic-model	13.1%	e,c,b,a
g	Topic-model + keyword	12.8%	f,e,c,b,a
h	Spectral + keyword	6.8%	d,a,b,c,e,f,g
i	Spectral + idiolectal	6.7%	h,d,a,b,c,e,f,g
j	Spectral + topic-model	6.2%	i,h,d,a,b,c,e,f,g
k	Spectral + idiolectal + topic-model	5.3%	j,i,h,d,a,b,c,e,f,g

Table 5.7: Equal error rates for the individual systems and their combinations.

The first noteworthy observation is that the idiolectal (a) and either of the topic-based systems (b, c) can be combined to reduce the error rate from 20.8% (a), 17.0% (b), and 16.3% (c) to 14.9% (e) and 13.1% (f), which can also be seen in Figure 5.12. Thus, it appears that although they all use features based on words, they indeed capture different and complementary types of speaker information. Also, the combination of both topic-based approaches (g) is significantly better at 12.8% EER than the individual systems (b, c), confirming that they concentrate on different kinds of information. It is likely that the topic-model-based approach is better at spotting the topic if an utterance contains topic marker words which have never been observed in the speaker's training data but which have been associated with the topic in the background material. On the other hand, the keyword-based method makes better use of names, titles, and other keywords that are associated much more with the person than with the topic.

The second noteworthy observation, which can also be seen in Figure 5.13, is that the performance of the voice-based (spectral) speaker recognition system (d) can indeed be improved by combination with the topic-based systems (h, j). This means that the topic-based features capture speaker information that is not captured by the spectral features, as is intuitively plausible. The fusion of the topic-model-based and the spectral system (j) outperforms both the keyword-based/spectral (h) and the idiolectal/spectral (i) fusions (6.2% EER versus 6.8% and 6.7%). A three-way combination of the topic-model-based, idiolectal, and spectral systems (k) further improves on this, reducing the EER to 5.3% by using voice, idiolect, and topic information.

It is also interesting to look at the individual systems' weights in the fusion to see which influence

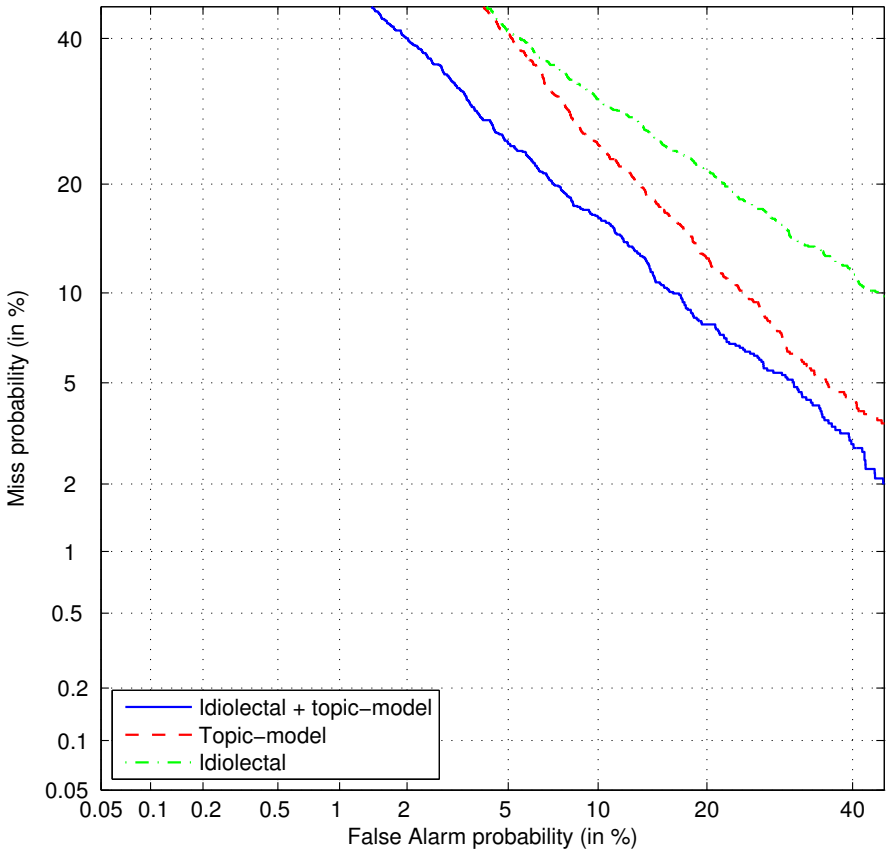


Figure 5.12: DET-curves for the topic-model and idiolectal systems and their combination.

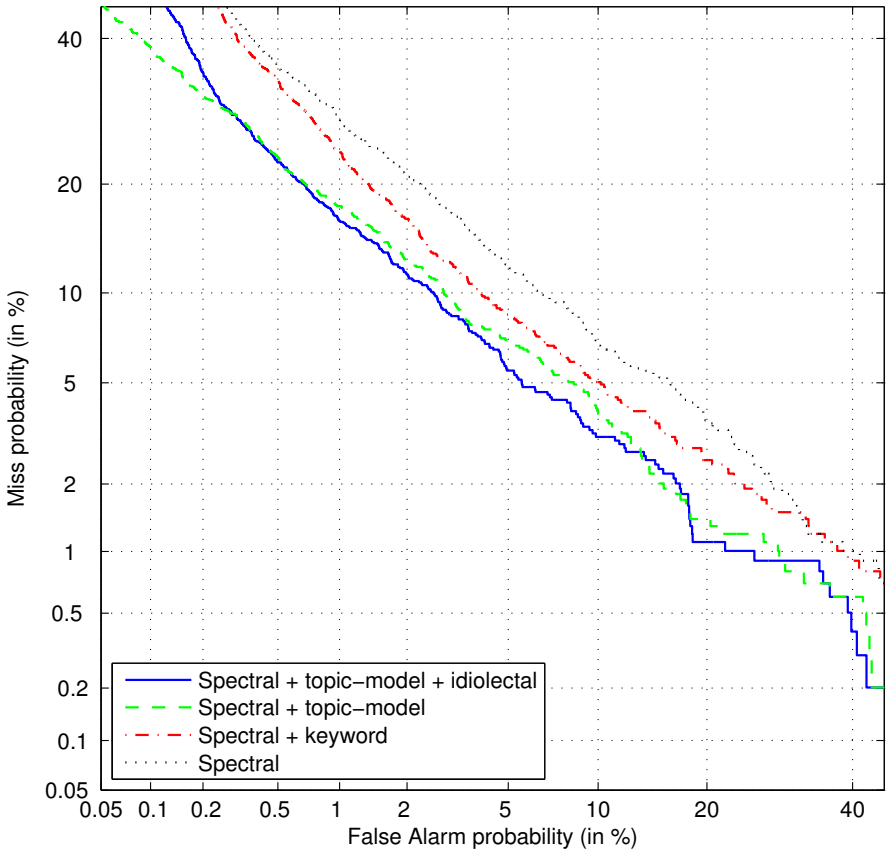


Figure 5.13: DET-curves for combining spectral, topic, and idiolectal systems.

they exert. To obtain directly comparable fusion weights, each system's scores where normalised to zero mean and standard variance before training the weights. The mean and standard deviation estimates for the normalisation were computed from the development scores (which were also used for weight training).

ID	System	Relative weights
e	Idiolectal + keyword	0.45 : 0.55
f	Idiolectal + topic-model	0.30 : 0.70
g	Topic-model + keyword	0.69 : 0.31
h	Spectral + keyword	0.70 : 0.30
i	Spectral + idiolectal	0.72 : 0.28
j	Spectral + topic-model	0.47 : 0.53
k	Spectral + idiolectal + topic-model	0.42 : 0.12 : 0.46

Table 5.8: Relative weights for the fused systems
(read: system A + system B | system A's weight : system B's weight).

Table 5.8 contains the relative weights of the subsystems in the system combinations from Table 5.7. The relative weights indicate how much the fused score s_{final} is influenced by the subsystems' scores in a simple linear combination ($s_{final} = w \cdot s_1 + (1-w) \cdot s_2$). For example, in their combination, the idiolectal and the keyword system have comparable weight (e), while in the pairwise combinations of topic-model-based system with idiolectal and keyword-based system (f, g), the former has a much stronger influence (about 70%) than the latter two (about 30% each). A similar ratio of 70% to 30% is found in the fusions of the spectral and the keyword-based and idiolectal systems (h, i), while the topic-model-based and the spectral system have comparable influence (j) in their fusion. This suggests that the spectral and topic-model-based systems bring more information to the fused system than the idiolectal and keyword-based systems. This is also indicated by the weights in the combination of spectral, idiolectal, and topic-model-based system (k). However, the information gained by adding a system to a fusion should in the end be judged by system performance measures, such as the equal error rates (Table 5.7) or DET-curves (Figures 5.9, 5.12, and 5.13), and not just from relative weights.

Finally, the usefulness of the test utterance length (number of content words) as side information in the fusion was tested. The FoCal Bilinear toolkit (which can make use of side information) was used for this (see Section 5.3). As the toolkit expects side information in the form of category membership probabilities for each trial, the content word counts were normalised to the interval $[0; 1]$ (by subtracting the smallest count and dividing by the difference of largest and smallest count). The normalised count can be seen as the trial's membership likelihood for the category "long utterances". Because each trial's probability distribution must sum to unity, the complementary probability $(1 - p)$ of the normalised count was used as a second category membership likelihood for category "short utterances".

The results of the system fusions with side information are listed in Table 5.9. The differences in equal

error rates between fusions with and without side information (i.e. in the same row) are all statistically significant according to McNemar's test at $\alpha = 0.01$. It appears that the two-way fusions of the spectral system with the keyword, idiolectal, and topic-model-based systems (h, i, j) can indeed be improved by adding the length as side information. However, in the three-way combination of spectral, idiolectal, and topic-model-based system (k), the performance worsens, so adding the length as side information cannot be recommended unequivocally.

ID	**System**	**EER without side information**	**EER with test utterance length as side information**
h	Spectral + keyword	6.8%	6.2%
i	Spectral + idiolectal	6.7%	6.4%
j	Spectral + topic-model	6.2%	6.0%
k	Spectral + idiolectal + topic-model	5.3%	5.5%

Table 5.9: Influence of test utterance length as side information in the fusion.

5.6.5 Analysis of Remaining Misclassifications

The previous section shows that the spectral speaker recognition system can be improved by fusing it with the idiolectal and the topic-model-based system, yielding a final equal error rate of 5.3%. To get an idea in which direction future research can be aimed, an analysis of the trials misclassified by that system was carried out.

There are 233,590 test trials (994 test files × 235 speakers) in the Bundestag corpus, of which 12,613 were misclassified by the spectral + topic-model + idiolectal fusion. As each test file is used in 235 trials, the "difficulty" of each test file can be assessed by its rate of erroneous trials. This error rate is computed by counting the test file's number of misclassified trials and dividing by 235. It turns out that there are a large number of "easy" test files which are hardly ever misclassified and a smaller number of "hard" test files with a high rate or erroneous trials. Figure 5.14 shows a histogram of the number of test files with particular misclassified trial rates. The easiest test files (with no or very few misclassifications) are most frequent, while the hard test files form a tail in the graph.

To gain some insight into what makes a test file easy or hard, the erroneous trial rates were correlated with some test file attributes: the recordings' overall duration, the duration of the speech segments in the files, the total word count, and the content word count. To see what influence the true speaker of a test file has, the average error rates for *other* test files by the same true speaker (referred to as "same speaker error rate") were used as an additional attribute. Pearson's correlation coefficient r was used to detect linear correlations between these attributes and the erroneous trial rates; r lies between -1 and 1,

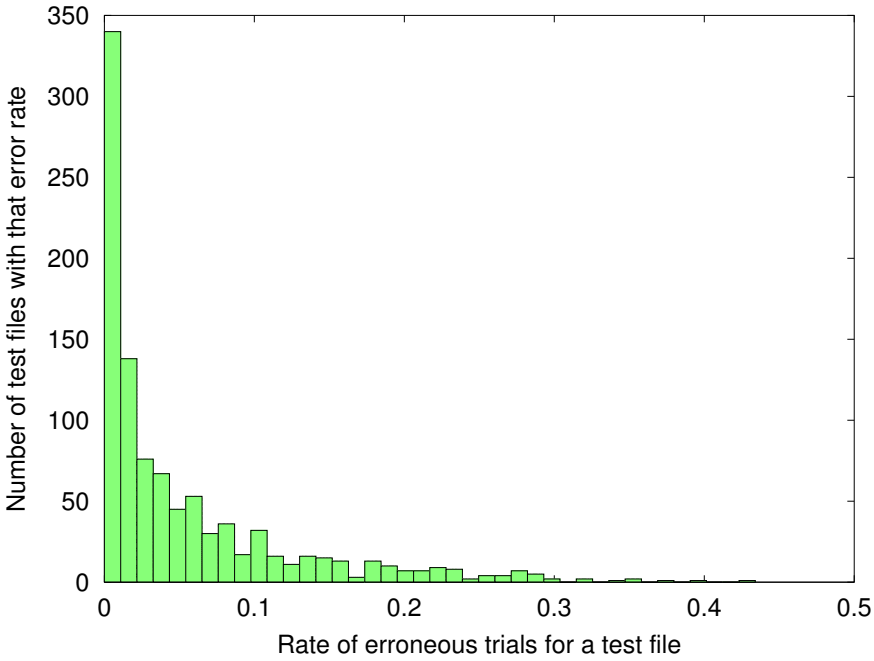


Figure 5.14: Histogram of the number of test files with particular trial error rates.

0 indicating independence, 1 perfect positive correlation, and -1 perfect negative correlation. The results are listed in Table 5.10.

	File duration	**Speech duration**	**Word count**	**Content word count**	**Same speaker error rate**
Correlation coefficient	0.12375	0.098303	0.11275	0.00092602	*0.54548*

Table 5.10: Correlation coefficients for correlation between the erroneous trial count and other attributes of a test file.

The results show, that longer duration (both overall and of the speech part) and higher word count of a test file do not correspond to fewer misclassifications for the fused system. This is not so surprising when looking at the amount of spoken content in the test files: 70% of them have at least 5 minutes and 93% have at least 2 minutes, which is sufficient to get good performance from the spectral and thus the fused system. The correlation might look very different if a test set with shorter files was employed. Also, the error rate seems mostly independent of the content word count (which may explain why using the content word count as side information in the fusion did not improve this system, see previous section). Most importantly, there is a strong correlation between the misclassifications rate of a test file and of other test files by the same speaker. This indicates that a few speakers are more difficult to recognise and easier to mistake for other speakers. This finding is reinforced when looking at the extreme ends of

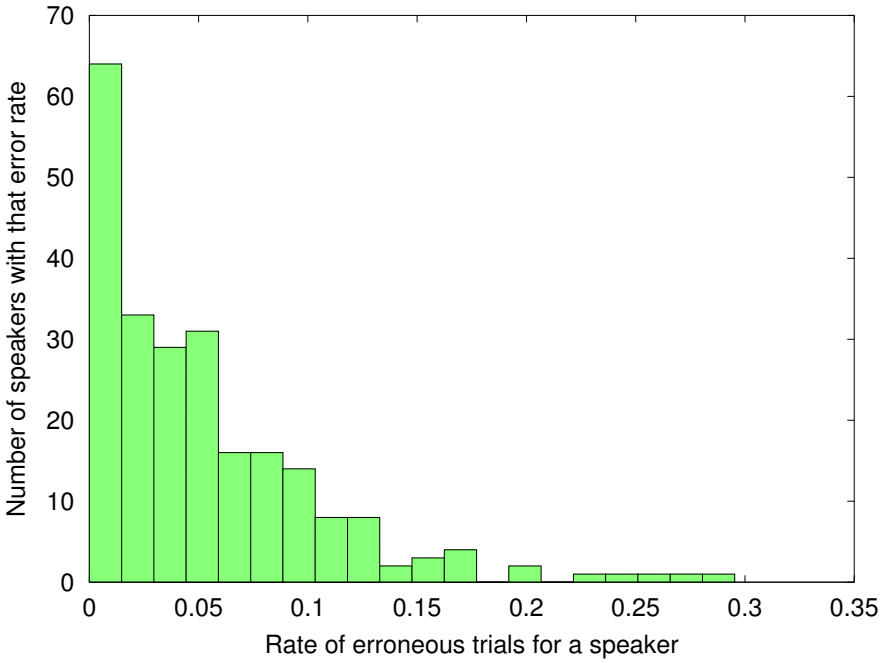


Figure 5.15: Histogram of the number of speakers with particular erroneous trial rates.

the error rate histogram: There are 142 test files from 86 speakers with no misclassifications at all, and 144 files from 80 speakers with an error rate of 12% or higher. Only 10 speakers have files in both sets. Empirical approximation shows that if the sets were randomly drawn (from the set of all test files), they should have been produced by 110 speakers each and 55 speakers should have test files in both sets. The histogram of erroneous trial rates per speaker in Figure 5.15 shows that many speakers are easy to classify but a small number is very hard to get right[8]. This effect seems to be common in speaker recognition systems: "Performance variability in speech and speaker recognition systems can be attributed to many factors. One major factor, which is often acknowledged but seldom analyzed, is inherent differences in the recognizability of different speakers." [Doddington98]

The effect cannot solely be caused by badly trained speaker models, because a speaker's model only influences the outcome of the true speaker trials, and there are at most 5 true speaker trials per speaker (each speaker has 2-5 test files) while the number of misclassified trials for hard speakers is much higher than 5. There may be a speaker characteristic which causes the effect, like a voice that is easy to confuse with others. It could also be caused by something not directly speaker-inherent which is nevertheless correlated with speaker identity. For example, some speakers may frequently cause heated discussion and noise in the background and thus decrease system performance on their test files. However, in a manual inspection of easy and hard files it was not immediately apparent what could cause the differences in

[8]The list of speaker error rates on the Bundestag test set is available online at `http://www.ifs.tuwien.ac.at/ir/speakerrec/`.

recognition performance. This could be a starting point for further research.

5.7 Influence of Speaker Interest Changes over Time

While the previous sections evaluated the usefulness of topic-based speaker recognition, its sensitivity to the utterances' lengths, and the improvement it brings to a combination of systems, the question how topic-based speaker recognition is influenced by topic and speaker interest changes will be investigated in this section.

There are two sources for change that may influence topic-based speaker recognition: First, the change of the topics themselves and second, the change of the speakers' interest in them. The topics may change by becoming more or less popular in the general public, and by new words being associated with them. For example, the public interest in energy politics surged in Germany after the meltdown at Fukushima, while at the same time the word "Fukushima" became a topic marker word for energy politics. On the other hand, specific speakers' interest in certain topics may change suddenly or over time, even if the topics themselves have existed for a long time. An intuitive example of that is the change of topic interests that occurs if a minister switches departments after a reorganisation of the cabinet.

These problems are commonly known as concept drift in machine learning. Various methods for handling concept drift in different application scenarios have been devised, for an overview see [Tsymbal04, Zliobaite09].

In particular, Dynamic Topic Models, similar to Latent Dirichlet Allocation, have been presented in [Blei06] and [Wang08], which explicitly deal with modelling the evolution of topics over time. In [Blei06] this is done by having one topic model with T topics for each time slice (for example one year) and chaining together the corresponding topics in the consecutive topic models. The natural parameters of each topic are chained in a state space model that evolves with Gaussian noise, so that the multinomial distribution $\phi_t^{(j)}$ over words for topic j at time t evolves from the distribution $\phi_{t-1}^{(j)}$. The method is extended in [Wang08] to continuous time dynamic topic models, which do not require time to be discretised into time slices because they replace the discrete state space model with its continuous generalisation, Brownian motion. Using Dynamic Topic Models, it should be possible to track the evolution of topics over time. Applying this to topic-based speaker recognition would be an interesting challenge for future work.

However, even if the evolution of topics is modelled, the speakers' interest changes are a second, unaddressed, source of concept drift. In order to measure how strong the influence of speaker interest changes on topic-based speaker recognition is, experiments were conducted. Unfortunately, the video

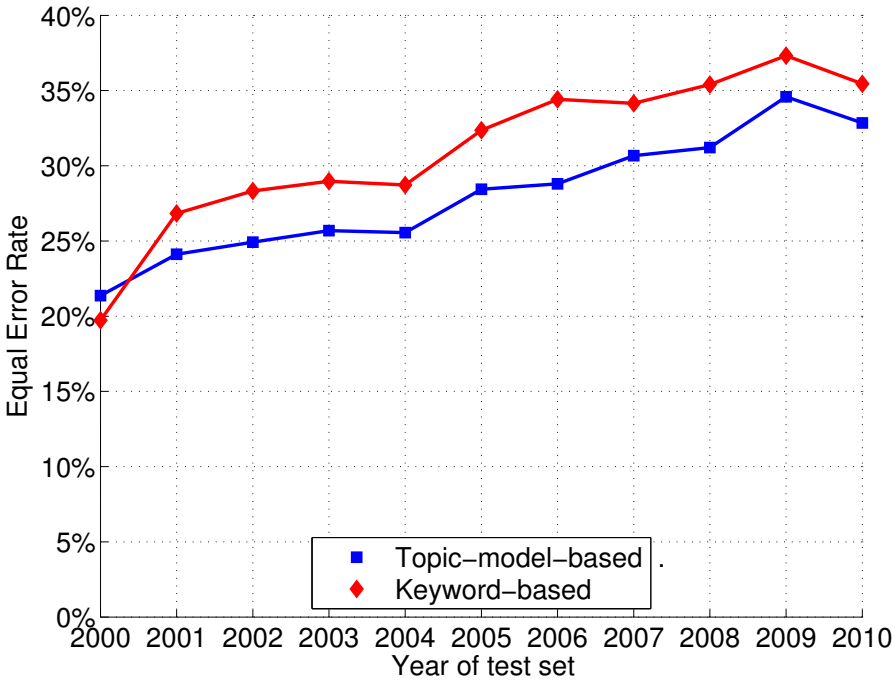


Figure 5.16: Influence of speaker interest change on topic-model-based speaker recognition. Models were trained on data from 1999.

recordings in the Bundestag corpus span only 16 months, so the long-term evolution of speaker interests could not be tracked in them. For this reason, a text-only corpus was developed from Bundestag stenographic transcripts, ranging from 1999 to 2010, which is described in Section 4.4.2. It contains speech transcripts from 97 speakers who were members of parliament during the whole examined timespan. This evaluation with written text is used as an approximation of the performance on audio data. Obviously, fusion experiments with the voice-based system cannot be done for the written text corpus.

Topic-based speaker recognition was tested on this set by training the speakers' models on data from 1999 and then running an evaluation with the data from each year from 2000 to 2010, obtaining equal error rates for each year. For the topic-model-based system, the topic model (with 20 topics) was trained on all data in the Speaker Preference Shift Evaluation Corpus (Section 4.4.2), in order to achieve the best possible topic model, so as to minimise the impact of topic drift and only investigate the influence of speaker interest drift. The topic mixtures for the target speakers' speeches were inferred with this topic model. Analogous to the keyword-based system, the speakers were modelled by prototype vectors using as training data the vectors for the 1999 set. For the keyword-based system, all speeches were converted to tf-idf vectors. Each speaker SVM was trained with the speaker's data from 1999 as the positive examples and all background speeches as negative examples. After training, both systems were tested on the data sets for 2000 to 2010. The resulting equal error rates are displayed in Figure 5.16 and Table 5.11

Test set year	Topic-model-based system	Keyword-based system
2000	21.4%	19.7%
2001	24.1%	26.8%
2002	24.9%	28.3%
2003	25.7%	29.0%
2004	25.6%	28.7%
2005	28.4%	32.4%
2006	28.8%	34.4%
2007	30.7%	34.2%
2008	31.2%	35.4%
2009	34.6%	37.3%
2010	32.9%	35.4%

Table 5.11: Equal error rates for both topic-based systems according to year of test set, trained on data from 1999.

The error rate gradually rises over the years as the time between training and test material becomes longer, as it would be expected. The keyword-based system performs worse than the topic-model-based system in most years, probably because it is more prone to degradation from topic drift and changing word preferences of speakers. It may overfit to keywords that were "en vogue" in 1999 because they were associated with current events but are out of use a few years later. On the other hand, the topic model used in the topic-model-based system maps words to the abstract topics discussed in the Bundestag, which are relatively constant even if some of the vocabulary to discuss them changes over the years. The topics can still be recognised by the more stable topic marker words. Thus the topic-model-based system seems to be more robust to changes in idiosyncratic word usage *if* the topic model can be trained with data from the whole timespan. It appears that in the Bundestag domain, speaker interest changes lead to a gradual increase of the equal error rate in the order of magnitude of 15% absolute over a decade.

To counter this effect, the speakers' models could be retrained regularly with new data to capture their current interests. This was simulated in two experiments: In the first, speaker models were trained with data from each single year and tested with the data from the following year. The results are shown in Figure 5.17. In the second experiment, all years before the test year was used for model training, the results are presented in Figure 5.18. Results for both experiments are also listed in Table 5.12.

Figure 5.17 confirms that by using training data from the previous year, the steady increase in error rate over time can be averted. Note that the EER fluctuation between the years is very likely caused by the varying sizes and number of speeches per speaker of the training and test sets (cf. Section 4.4.2). An ideal corpus for this experiment would have the same number of speeches for each speaker in each year – and a high number of speakers – but not enough real-life data was available to construct such a uniform corpus.

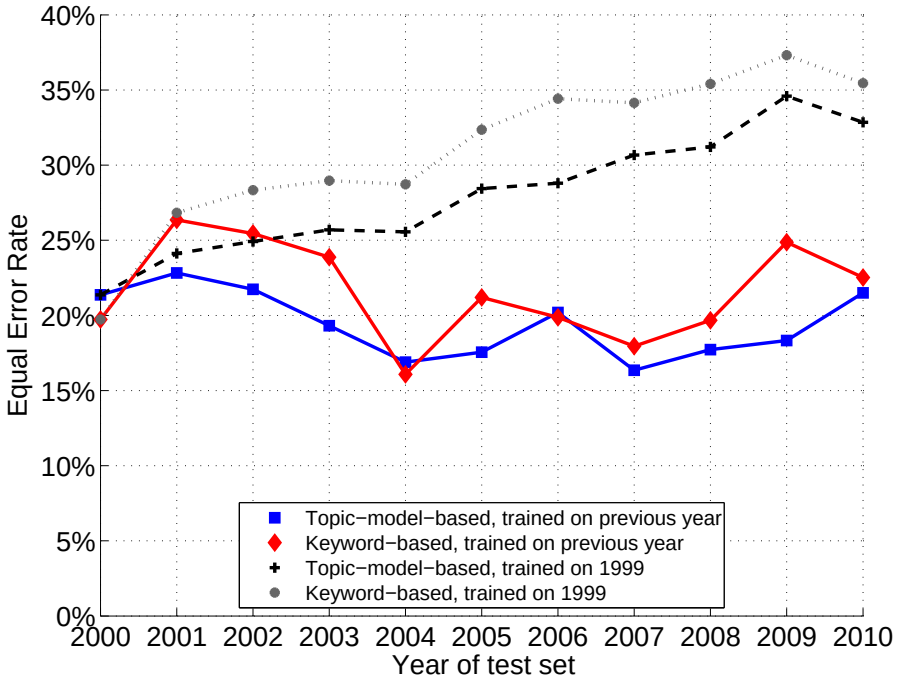


Figure 5.17: Equal error rates for each test set year when models are trained on data from the previous year.

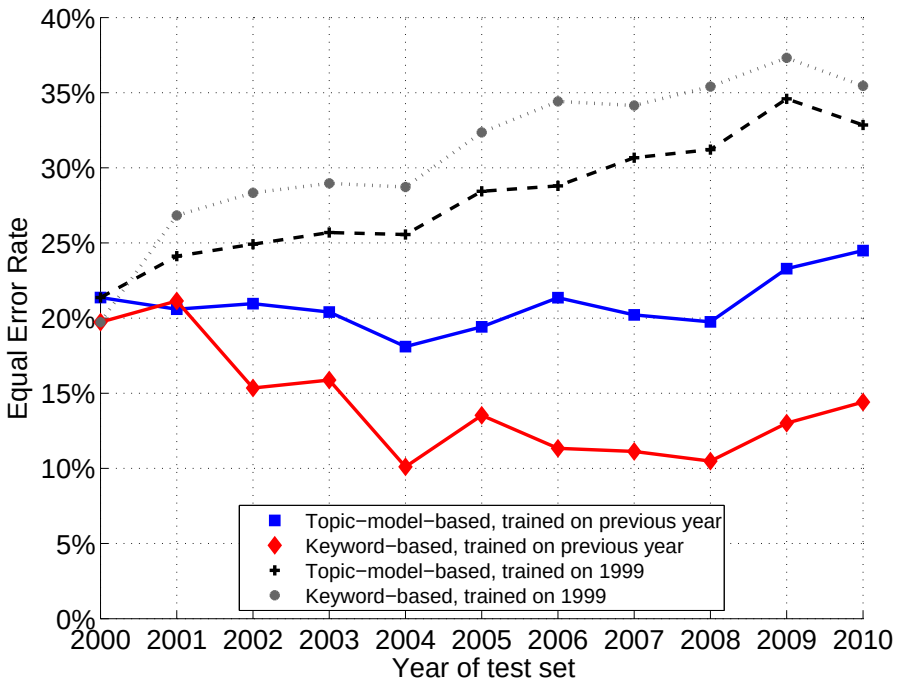


Figure 5.18: Equal error rates for each test set year when models are trained on data from *all* previous years since 1999.

Test set year	Previous year		All previous years	
	Topic-model-based system	Keyword-based system	Topic-model-based system	Keyword-based system
2000	21.4%	19.7%	21.4%	19.7%
2001	22.8%	26.4%	20.6%	21.1%
2002	21.7%	25.4%	21.0%	15.3%
2003	19.3%	23.9%	20.4%	15.9%
2004	16.9%	16.1%	18.1%	10.1%
2005	17.6%	21.2%	19.4%	13.5%
2006	20.2%	19.9%	21.4%	11.3%
2007	16.3%	17.9%	20.2%	11.1%
2008	17.7%	19.7%	19.7%	10.5%
2009	18.3%	24.9%	23.3%	13.0%
2010	21.5%	22.5%	24.5%	14.4%

Table 5.12: Equal error rates for both topic-based systems according to year of test set, trained on data from the previous year and from all previous years.

Figure 5.18 shows that the keyword-based system benefits considerably from additional training data from all years before the test set year, while the topic-model-based system does not. This is very likely caused by the different speaker modelling methods used in both approaches: The more complex SVM models of the keyword-based system can take advantage of the higher number of training vectors, while the very simple prototype vector models cannot capture the additional information. The experiments in this thesis assume a usage scenario where only a few training examples per speaker are available. In a different scenario where more examples are available, a different modelling method should be considered for a topic-model-based system.

So one possible approach to counter error rate increases due to speaker topic drift is to re-train speaker models with examples close in time to the test data. However, as obtaining additional training data is costly and time-consuming if not impossible for many real-life scenarios, modelling topic drift and finding other practical methods to deal with speaker interest changes are an interesting field of future work.

5.8 Summary

This chapter presents the evaluations carried out to explore topic-based speaker recognition: The methods described in Chapter 3 were implemented using a combination of tools like the Snowball stemmer, the Mallet topic modelling toolkit, the SVMlight Support Vector Machine libraries, and custom code. Also, two baseline systems were set up, one using voice-based and one with idiolectal features. During system calibration, the SVM-based speaker modelling method was selected for the keyword-based system because it yielded a lower error rate. Also, it was confirmed that utterance pre-processing has a consid-

erable impact on topic-based speaker recognition: stemming and stop word removal are crucial for good performance, even if this should sacrifice information about the speakers' style contained in the function words. At least for the Bundestag data, common tf-idf weighting seems to outperform rank normalisation as a weighting scheme for keyword-based and idiolectal speaker recognition. The Bundestag data can be represented well for speaker recognition purposes by just 20 topics. For the topic-model-based system, using Kullback-Leibler divergence for utterance scoring gave the best performance. Also, the simple averaging of all training examples' topic vectors into a prototype vector gives performance superior to SVM modelling, presumably because of the very few training examples.

The main evaluation indicates that the topics found by unsupervised topic modelling in the Bundestag background data make sense to humans, and that well-known speakers are assigned their expected topic preferences. In comparison to the baseline systems, topic-based speaker recognition (topic-model-based: EER 16.3%, keyword-based: EER 17.0%) seems to outperform the idiolectal system (EER 20.8%) but to be less reliable than the voice-based approach (EER 8.6%). It was shown that the length of the test utterance influences the reliability of topic-based speaker recognition systems, as was expected. For utterances with less than 100 content words, error rates increase dramatically, while for longer utterances good results are achieved. The information captured in the topic-based systems' scores can be successfully fused with the other systems to increase performance. For example, a combination of spectral and topic-model-based system improves the EER from 8.6% to 6.2%, while a three-way combination of the topic, idiolectal, and spectral systems reduces the EER to 5.3%. This shows that topic preferences are a speaker recognition cue complementary to the others, and that their use can be very beneficial for a system. Both topic-based approaches, the keyword-based and the topic-model-based method, give comparable results, with the topic-model-based system being slightly better (EER 16.3% vs. 17.0%). This is interesting because of the fundamentally different feature vectors employed by both systems: The keyword-based approach uses high-dimensional (approx. 61,000 dimensions) but sparse tf-idf vectors and models the speakers' preferred words, while the topic-model-based approach uses low-dimensional (20 dimensions) topic mixture vectors and models the speakers' preferred topics directly. Thus, the topic mixture vectors are a more compact way of representing the topic information that is obviously used by both systems. Also, using side information about the test utterance length in the fusion can lead to lower error rates, but this is not always the case. The misclassifications made by the best fused system (the combination of spectral, idiolectal, and topic-model-based systems) were analysed to see which characteristics would lead to lower performance. It seems that there is some speaker-related characteristic which leads to a higher error rate and which should be addressed in future research.

In an evaluation on a written text dataset spanning a whole decade of political speeches from 1999 to 2010 in parliament, the influence of changes in the speakers' topic interests was estimated. Models

were trained on data from 1999 and a test set for each following year was scored against them. Over the whole decade, speaker interest drift caused an increase in equal error rates in the order of magnitude of 15% absolute. It was found that this performance decrease can be averted by retraining the speaker models with data closer in time to the test data: Models trained on data from the previous year for each yearly test set exhibited a much lower error rate than the original models from 1999.

Chapter 6

Conclusion and Future Work

In this thesis, it was investigated in how far topic information can be used to aid automatic speaker recognition. Reliable speaker recognition methods are required to automatically annotate multimedia archives, making audio and video recordings searchable by speaker. In areas such as the broadcast domain, it can be expected that the topic of an utterance or an utterance's context carries information about the likely speakers, which can be used to improve upon traditional voice-based speaker identification.

The first research question in this work was how to automatically capture spoken utterance's topic information with little manual effort, and how to use it to learn speakers' topic preferences (Q1, see Section 1.1). Two methods to automatically learn speaker topic preferences from training audio data have been presented, one based on implicit and one based on explicit topic representation. In both cases, automatic speech recognition is used to obtain a word transcript from the audio material, as the topic information is carried mostly by the words in an utterance. Text pre-processing is applied to condense the topic information in the transcript. For implicit representation, the speakers' topic preferences are modelled by learning their idiosyncratic words, which will consist mostly of topic marker words for their preferred topics. Two modelling methods were detailed: Log-likelihood ratio scoring, which compares word frequencies in speaker and background data, and Support Vector Machines trained on tf-idf vectors. It was found that the SVM method gives slightly better performance. For explicit topic representation, unsupervised probabilistic topic modelling can be applied to a text collection from the domain to automatically find the prevalent topics and identify the corresponding marker words. Using an unsupervised method removes the need for labelled topic training data or even knowing what the prevalent topics are. With the resulting topic model, a topic mixture can be inferred for each speaker utterance, reducing utterances to topic feature vectors. These vectors can be used in various speaker modelling schemes, the most successful and simple scheme in this thesis was found to be representing

speaker interests by a topic prototype vector.

The second research question was how topic-based speaker recognition performs and whether it combines well with other cues, such as voice-based speaker recognition (Q2). As no benchmark corpus for speaker recognition was available that allows testing of topic-based speaker recognition (because in the standard corpora speakers cannot chose their topics freely), a new corpus was built to enable a thorough evaluation. It contains 324 hours of video recordings of speeches given in German parliament (the Bundestag) by 235 different politicians, who can be expected to talk about their departments and fields of expertise. It also has a text collection from the domain to train the topic model on. In the evaluation it was found that both proposed topic and speaker modelling approaches can be successfully used for speaker recognition. The obtained equal error rates of 17.0% for the implicit, keyword-based and 16.3% for the explicit, topic-model-based approach were not as low as for a standard voice-based reference system (EER 8.6%), but lower than for an idiolectal baseline system (EER 20.8%). More importantly, the scores from the topic-based and spectral systems could be fused to improve on the spectral system's performance, which means that topic profiles are complementary to voice characteristics. Combining the topic-model-based and the spectral system decreased the EER from 8.6% to 6.2%, a three-way combination of topic-model-based, spectral, and idiolectal system further reduced the EER to 5.3%. Both topic-based methods perform comparably (at EERs 17.0% and 16.3%), which means that the topic information contained in 61,000-dimensional sparse tf-idf vectors can also be captured in only 20-dimensional topic mixture vectors for speaker recognition purposes. It was estimated that an utterance should have at least 100 content words to allow reliable topic-based speaker identification with the described modelling methods. Also, in some cases side-information about the number of content words in the test utterance can be used to improve the fusion of voice-based and topic-based systems.

The third research question was how stable the speakers' topic preferences are over time (Q3). To gauge the influence of speaker preferences shift, an extension to the Bundestag corpus was put together from stenographic transcripts covering the timespan from 1999 to 2010. It has 4,369 utterances from 97 target speakers who where members of parliament during that timespan, and 33,571 background utterances from other speakers. The topic-model-based and keyword-based systems were evaluated on this extension corpus, and it was found that the changes of the speakers' topic interests degrade topic-based speaker recognition. The equal error rate of models trained on data from 1999 rose from around 20% to around 35% over 10 years. Retraining speaker models annually with data from previous years averted this problem. It also showed that the keyword-based system benefits strongly from more training data, probably because of the more complex SVM modelling. If it is to be expected that more training data is available in a domain, the topic-model-based system should also be improved with a more complex modelling scheme.

This thesis constitutes a first in depth study of using topic as a cue in speaker recognition, as topic is still largely disregarded in the existing speaker recognition literature. It addresses the fundamental research questions regarding topic cues, but it cannot answer all conceivable questions about topic-based speaker recognition, so numerous possibilities for future research remain. For example, it would be interesting to quantitatively evaluate in how far the automatically derived topics and speaker/topic assignments make sense and are intuitive to humans. Human subjects could be asked whether they agree with a speaker's automatically derived topic assignments.

As the automatic estimation of speaker topic preferences requires more training material than voice-based speaker recognition, ways to reduce the need for training data or at least the cost for obtaining it would be worthwhile investigating. Could manually estimated topic preferences be successfully used for automatic speaker recognition? With appropriately designed questionnaires, human subjects' intuition about speakers' preferred topics could be transformed into topic mixture vectors and evaluated as models in automatic speaker recognition. Also, it may be possible to augment or replace audio training data with written text automatically retrieved from the Internet: Search engines could be queried with the speaker's name, the results should be filtered and suitable pages could be retrieved, processed and used as text training material.

Especially with more training data, a number of alterations, improvements, and extensions are conceivable for topic-based speaker recognition. Speaker modelling in the topic-model-based approach could be refined to something more complex than one speaker vector prototype. For example, SVM models, with specialised or standard kernels, could be re-evaluated with more training data. It would also be interesting to investigate whether speakers can be modelled in a way that makes use of the chronological order of training examples – which may be a way to handle the concept drift stemming from speaker interest changes.

The topic-modelling process is also a candidate for refinement: Latent Dirichlet Allocation has been extended in various directions, and it would be interesting to test whether these extensions benefit topic-based speaker recognition. For example, the use of dynamic topic models [Blei06, Wang08] to handle topic concept drift over time could be evaluated. Also, in [Branavan08], the authors adapt the LDA model to fit text documents annotated with tags and key-phrases. A document's topic mixture is then derived from both its text and its associated annotations (if present). In case there are no, too few, or noisy annotations, the topic can still be derived from the text. Integrating annotations in this way can make topic identification more robust without relying solely on them, so that documents without annotations can still be processed. In the case of speeches from sessions in parliament, the key-phrases could be taken from the written agendas, even if they give no exact timestamps for when in the session which of the items is discussed.

Other LDA extensions were proposed for documents containing images, e.g. in [Feng10] and [Blei03a], where document and image "terms" are jointly modelled. Both approaches use (different) image region features as visual terms, which make up a visual vocabulary that can be used to describe (and model) images. The visual vocabulary is treated as an extension of the word vocabulary. Accordingly, the probability distribution that represents a topic covers both visual terms and words. Incorporating image information into topic determination for speaker recognition may benefit application domains where video is available (in addition to audio); of course, in order to be useful, the video should contain topic cues. This would likely be the case e.g. for broadcast news containing footage of the reported incident.

Finally, due to the structure of the audio files available from German parliament when the Bundestag corpus was set up, the experiments in this thesis were carried out on recordings of individual speeches containing only one speaker. In the meantime, recordings from committee meetings are available online, thus topic-based speaker recognition could be tested on material containing several speakers, given a speaker diarization system. Usually, in the committee sessions several politicians talk about the same topic on which they are considered experts. The information about the topic of the discussion would significantly narrow down the list of potential speakers, in this case from hundreds of politicians to probably under twenty. The discrimination between speakers on the same topic could then be done by spectral characteristics (voice).

A lot of other interesting questions can be conceived, and the suggestions in this chapter cannot be exhaustive. As it was shown that speaker recognition systems indeed benefit from incorporating topic cues, it can be expected that investigating the open questions would lead to further improvements.

Bibliography

[Abbasi05] *Ahmed Abbasi and Hsinchun Chen.* "Applying authorship analysis to extremist-group web forum messages." *IEEE Intelligent Systems*, volume 20:pp. 67–75, September 2005.

[Andrews01] *Walter D. Andrews, Mary A. Kohler, and Joseph P. Campbell.* "Phonetic speaker recognition." In *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pp. 2517–2520. Aalborg, Denmark, September 2001.

[Andrews02] *Walter D. Andrews, Mary A. Kohler, Joseph P. Campbell, John J. Godfrey, and Jaime Hernandez-Cordero.* "Gender-dependent phonetic refraction for speaker recognition." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 (ICASSP '02)*, volume 1, pp. 149–152. 2002.

[Argamon05] *Shlomo Argamon and Shlomo Levitan.* "Measuring the usefulness of function words for authorship attribution." In *Proceedings of Association for Computers and the Humanities (ACH) / Association for Literary and Linguistic Computing (ALLC) Joint Conference 2005*, pp. 4–6. Victoria, Canada, June 2005.

[Atal69] *Bishnu Saroop Atal.* "Automatic speaker recognition based on pitch contours." *The Journal of the Acoustical Society of America*, volume 45, no. 1:p. 309, January 1969.

[Auckenthaler00] *Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas.* "Score normalization for text-independent speaker verification systems." *Digital Signal Processing*, volume 10, no. 1–3:pp. 42–54, 2000.

[Baker04] *Brendan Baker, Robbie Vogt, Michael Mason, and Sridha Sridharan.* "Improved phonetic and lexical speaker recognition through MAP adaptation." In *ODYSSEY04 - The Speaker and Language Recognition Workshop*, pp. 91–96. Toledo, Spain, May 2004.

[Baum09a] *Doris Baum.* "Topic-based speaker recognition for german parliamentary speeches." In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '09)*, pp. 427–431. Merano, Italy, December 2009.

[Baum09b] *Doris Baum, Barbara Samlowski, Thomas Winkler, Rolf Bardeli, and Daniel Schneider.* "DiSCo - a speaker and speech recognition evaluation corpus for challenging problems in the broadcast domain." In *Wolfgang Hoeppner* (editor), *Proceedings of the GSCL Symposium 'Sprachtechnologie und eHumanities' 2009*, pp. 1–9. Duisburg, Germany, February 2009.

[Baum10a] *Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler.* "DiSCo — a german evaluation corpus for challenging problems in the broadcast domain." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pp. 1695–1699. Malta, 2010.

[Baum10b] *Doris Baum, Daniel Schneider, Timo Mertens, and Joachim Köhler.* "Constrained subword units for speaker recognition." In *Odyssey 2010, The Speaker and Language Recognition Workshop*, pp. 1–6. Brno, Czech Republic, 2010.

[Baum12] *Doris Baum.* "Recognising speakers from the topics they talk about." *Speech Communication*, volume 54, no. 10:pp. 1132–1142, 2012.

[Biatov02] *Konstantin Biatov and Joachim Köhler.* "Methods and tools for speech data acquisition exploiting a database of German parliamentary speeches and transcripts from the internet." In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC '02)*, pp. 883–886. Las Palmas, Spain, June 2002.

[Bishop95] *Christopher M. Bishop. Neural Networks for Pattern Recognition.* Oxford University Press, 1 edition, January 1995.

[Blei03a] *David M. Blei and Michael I. Jordan.* "Modeling annotated data." In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03, pp. 127–134. Toronto, Canada, 2003.

[Blei03b] *David M. Blei, Andrew Y. Ng, and Michael I. Jordan.* "Latent dirichlet allocation." *Journal of Machine Learning Research*, volume 3:pp. 993–1022, 2003.

[Blei06] *David M. Blei and John D. Lafferty.* "Dynamic topic models." In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 113–120. Pittsburgh, USA, 2006.

[Boersma93] *Paul Boersma.* "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." In *Institute of Phonetic Sciences, University of Amsterdam, Proceedings*, volume 17, pp. 97–110. Amsterdam, The Netherlands, 1993.

[Branavan08] *Satchuthananthavale R.K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay.* "Learning document-level semantic properties from free-text annotations." In *Proceedings the 46th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies (ACL-08: HLT)*, pp. 263–271. Association for Computational Linguistics, Columbus, Ohio, USA, June 2008.

[Brümmer06] *Niko Brümmer and Johan du Preez.* "Application-independent evaluation of speaker detection." *Computer Speech & Language*, volume 20, no. 2-3:pp. 230–275, April 2006.

[Brümmer07] *Niko Brümmer, Lukáš Burget, Jan Černocký, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A. van Leeuwen, Pavel Matějka, Petr Schwarz, and Albert Strasheim.* "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, no. 7:pp. 2072–2084, 2007.

[Brümmer10] *Niko Brümmer, Lukáš Burget, Patrick Kenny, Pavel Matějka, Edward de Villiers, Martin Karafiát, Marcel Kockmann, Ondrej Glembek, Oldrich Plchot, Doris Baum, and Mohammed Senoussauoi.* "ABC system description for NIST SRE 2010." In *Proceedings of the NIST 2010 Speaker Recognition Evaluation*, pp. 1–20. VUT v Brno, Brno, Czech Republic, 2010.

[Burget07] *Lukáš Burget, Pavel Matějka, Petr Schwarz, Ondrej Glembek, and Jan Černocký.* "Analysis of feature extraction and channel compensation in a GMM speaker recognition system." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, no. 7:pp. 1979–1986, 2007.

[Burget09] *Lukáš Burget, Pavel Matějka, Valiantsina Hubeika, and Jan Černocký.* "Investigation into variants of joint factor analysis for speaker recognition." In *10th Annual Conference of the International Speech Communication Association (Interspeech '09)*, 9, pp. 1263–1266. International Speech Communication Association, 2009.

[Burrows02] *John Burrows.* "Delta: a measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing*, volume 17, no. 3:pp. 267–287, 2002.

[Campbell03a] *Joseph P. Campbell, Douglas A. Reynolds, and Robert B. Dunn.* "Fusing high- and low-level features for speaker recognition." In *Eurospeech 2003*, pp. 2665–2668. Geneva, Switzerland, 2003.

[Campbell03b] *William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, Douglas A. Jones, and Timothy R. Leek.* "Phonetic speaker recognition with support vector machines." In *Neural Information Processing Systems Conference*, pp. 1377–1384. Vancouver, Canada, December 2003.

[Campbell06a] *William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, E. Singer, and Pedro A. Torres-Carrasquillo.* "Support vector machines for speaker and language recognition." *Computer Speech & Language*, volume 20, no. 2-3:pp. 210–229, April 2006.

[Campbell06b] *William M. Campbell, D.E. Sturim, and Douglas A. Reynolds.* "Support vector machines using gmm supervectors for speaker verification." *Signal Processing Letters, IEEE*, volume 13, no. 5:pp. 308–311, May 2006.

[Canseco-Rodriguez04] *Leornardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain.* "Speaker diarization from speech transcripts." In *International Conference on Speech and Language Processing*, pp. 1272–1275. Jeju Island, October 2004.

[Canseco05] *Leonardo Canseco, Lori Lamel, and Jean-Luc Gauvain.* "A comparative study using manual and automatic transcriptions for diarization." In *Proceedings of IEEE Workshop on Automatic Speech Recognition.* San Juan, Porto Rico, November 2005.

[Carey96] *Micheal J. Carey, Eluned S. Parris, Harvey Lloyd-Thomas, and Stephen Bennett.* "Robust prosodic features for speaker identification." In *Fourth International*

Conference on Spoken Language 1996 (ICSLP 96), volume 3, pp. 1800–1803. Philadelphia, USA, 1996.

[Champod00] *Christophe Champod and Didier Meuwly.* "The inference of identity in forensic speaker recognition." *Speech Communication*, volume 31, no. 2-3:pp. 193–203, June 2000.

[Chang09] *Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei.* "Reading tea leaves: How humans interpret topic models." In *Neural Information Processing Systems*, volume 31, pp. 1–9. Vancouver, Canada, 2009.

[Chaski05] *Carole E. Chaski.* "Who's at the keyboard: Authorship attribution in digital evidence investigations." *International Journal of Digital Evidence*, volume 4, no. 1:pp. 1–13, 2005.

[Cheng10] *Shih-Sian Cheng, Hsin-Min Wang, and Hsin-Chia Fu.* "BIC-Based speaker segmentation using Divide-and-Conquer strategies with application to speaker diarization." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 18, no. 1:pp. 141–157, 2010.

[Cheveigné02] *Alain de Cheveigné and Hideki Kawahara.* "YIN, a fundamental frequency estimator for speech and music." *The Journal of the Acoustical Society of America*, volume 111, no. 4:pp. 1917–1930, 2002.

[Cieri04a] *Christopher Cieri, Joseph P. Campbell, Hirotaka Nakasone, David Miller, and Kevin Walker.* "The mixer corpus of multilingual, multichannel speaker recognition data." In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, pp. 26–28. Lisbon, Portugal, 2004.

[Cieri04b] *Christopher Cieri, David Miller, and Kevin Walker.* "The fisher corpus: a resource for the next generations of speech-to-text." In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, pp. 69–71. Lisbon, Portugal, 2004.

[Cortes95] *Corinna Cortes and Vladimir Vapnik.* "Support-vector networks." *Machine Learning*, volume 20, no. 3:pp. 273–297, 1995.

[Davis80] *Steven B. Davis and Paul Mermelstein.* "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In

IEEE Transactions on Acoustics, Speech and Signal Processing, volume 28, pp. 357–366. August 1980.

[Dehak07] *Najim Dehak, Pierre Dumouchel, and Patrick Kenny.* "Modeling prosodic features with joint factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing*, volume 15, no. 7:pp. 2095–2103, 2007.

[Diederich03] *Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass.* "Authorship attribution with support vector machines." *Applied Intelligence*, volume 19:pp. 109–123, 2003.

[Doddington98] *George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds.* "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation." In *5th International Conference on Spoken Language Processing (ICSLP'98)*, pp. 1351–1355. Sydney, Australia, December 1998.

[Doddington00] *George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds.* "The nist speaker recognition evaluation - overview, methodology, systems, results, perspective." *Speech Communication*, volume 31, no. 2–3:pp. 225–254, 2000.

[Doddington01] *George R. Doddington.* "Speaker recognition based on idiolectal differences between speakers." In *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pp. 2521–2524. Aalborg, Denmark, 2001.

[Dumais98] *Susan Dumais, John Platt, David Heckerman, and Mehran Sahami.* "Inductive learning algorithms and representations for text categorization." In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM '98)*, pp. 148–155. Bethesda, USA, 1998.

[Feng10] *Yansong Feng and Mirella Lapata.* "Topic models for image annotation and text illustration." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 831–839. Association for Computational Linguistics, Los Angeles, California, June 2010.

[Ferrer08] *Luciana Ferrer, Martin Graciarena, Argyris Zymnis, and Elizabeth Shriberg.* "System combination using auxiliary information for speaker verification." In

IEEE International Conference on Acoustics, Speech and Signal Processing 2008 (ICASSP '08), pp. 4853–4856. IEEE, Las Vegas, USA, April 2008.

[Ferrer10a] *Luciana Ferrer, Martin Gracianrena, Sachin Kajarekar, Nicolas Scheffer, Elizabeth Shriberg, and Andreas Stolcke.* "2010 NIST speaker recognition evaluation: SRI system description." In *NIST 2010 Speaker Recognition Evaluation.* Brno, Czech Republic, 2010.

[Ferrer10b] *Luciana Ferrer, Nicolas Scheffer, and Elizabeth Shriberg.* "A comparison of approaches for modeling prosodic features in speaker recognition." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 4414–4417. Dallas, Texas, USA, 2010.

[Galliano06] *Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-Francois Bonastre, Djamel Mostefa, and Khalid Choukri.* "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pp. 139–142. 2006.

[Gillick89] *Larry Gillick and Stephen J. Cox.* "Some statistical issues in the comparison of speech recognition algorithms." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 1989 (ICASSP-89)*, volume 1, pp. 532 –535. May 1989.

[Goldstein-Stewart09] *Jade Goldstein-Stewart, Ransom Winder, and Roberta Evans Sabin.* "Person identification from text and speech genre samples." In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pp. 336–344. Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

[Goldwater10] *Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning.* "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates." *Speech Communication*, volume 52, no. 3:pp. 181–200, 2010.

[Griffiths04] *Thomas L. Griffiths and Mark Steyvers.* "Finding scientific topics." *Proceedings of the National Academy of Sciences*, volume 101, no. Suppl 1:pp. 5228–5235, 2004.

[Hatch05] *Andrew O. Hatch, Barbara Peskin, and Andreas Stolcke.* "Improved phonetic speaker recognition using lattice decoding." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 (ICASSP '05)*, volume 1, pp. 169–172. March 2005.

[Hermansky90] *Hynek Hermansky.* "Perceptual linear predictive (PLP) analysis of speech." *The Journal of the Acoustical Society of America*, volume 87, no. 4:pp. 1738–1752, April 1990.

[Hsu03] *Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin.* "A practical guide to support vector classification." http://www.csie.ntu.edu.tw/˜cjlin/papers/guide/guide.pdf, 2003. Tutorial.

[Jin03] *Qin Jin, Jiri Navratil, Douglas A. Reynolds, Joseph P. Campbell, Walter D. Andrews, and J.S. Abramson.* "Combining cross-stream and time dimensions in phonetic speaker recognition." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, volume 4, pp. 800–803. Hong Kong, April 2003.

[Joachims98] *Thorsten Joachims.* "Text categorization with support vector machines: Learning with many relevant features." In *Proceedings of the 10th European Conference on Machine Learning.* Springer, Chemnitz, Germany, 1998.

[Joachims99] *Thorsten Joachims.* "Making large-scale SVM learning practical." In *B. Schölkopf, C. Burges, and A. Smola* (editors), *Advances in Kernel Methods - Support Vector Learning*, chapter 11. MIT Press, Cambridge, MA, 1999.

[Kenny04] *Patrick Kenny and Pierre Dumouchel.* "Disentangling speaker and channel effects in speaker verification." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 1, pp. 37–40. Montreal, Canada, 2004.

[Kenny06] *Patrick Kenny.* "Joint factor analysis of speaker and session variability : Theory and algorithms." Technical report, CRIM Montreal, 2006.

[Kenny08] *Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel.* "A study of inter-speaker variability in speaker verification." *IEEE Transactions on Audio Speech and Language Processing*, volume 16:pp. 980–988, 2008.

[Kinnunen10] *Tomi Kinnunen and Haizhou Li.* "An overview of text-independent speaker recognition: From features to supervectors." *Speech Communication*, volume 52, no. 1:pp. 12–40, 2010.

[Klusacek03] *David Klusacek, Jiri Navratil, Douglas A. Reynolds, and Joseph P. Campbell.* "Conditional pronunciation modeling in speaker detection." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, volume 4, pp. 804–807. April 2003.

[Kockmann08] *Marcel Kockmann and Lukáš Burget.* "Contour modeling of prosodic and acoustic features for speaker recognition." In *IEEE Spoken Language Technology Workshop (SLT 2008)*, pp. 45–48. Goa, India, 2008.

[Kockmann10] *Marcel Kockmann, Lukáš Burget, and Jan Černocký.* "Investigations into prosodic syllable contour features for speaker recognition." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, volume 2010, pp. 4418–4421. Dallas, US, 2010.

[Lee01] *A. Lee, T. Kawahara, and K. Shikano.* "Julius — an open source real-time large vocabulary recognition engine." In *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pp. 1691–1694. ISCA, Aalborg, Denmark, September 2001.

[Martin97] *Alvin Martin, George R. Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki.* "The DET curve in assessment of detection task performance." In *Eurospeech '97*, pp. 1895–1898. Rhodes, Greece, 1997.

[Mauclair06] *Julie Mauclair, Sylvain Meignier, and Yannick Esteve.* "Speaker diarization: About whom the speaker is talking?" In *IEEE Speaker and Language Recognition Workshop 2006 (Odyssey 2006)*, pp. 1–6. San Juan, Puerto Rico, June 2006.

[McCallum02] *Andrew Kachites McCallum.* "Mallet: A machine learning for language toolkit." http://mallet.cs.umass.edu, 2002. Software toolbox.

[McNemar47] *Quinn McNemar.* "Note on the sampling error of the difference between correlated proportions or percentages." *Psychometrika*, volume 12:pp. 153–157, 1947.

[Mertens09] *Timo Mertens and Daniel Schneider.* "Efficient subword lattice retrieval for german spoken term detection." In *IEEE International Conference on Acous-*

tics, Speech and Signal Processing, ICASSP 2009, pp. 4885–4888. IEEE, Taipei, Taiwan, April 2009.

[Miyajima01] *Chiyomi Miyajima, Hideyuki Watanabe, Keiichi Tokuda, Tadashi Kitamura, and Shigeru Katagiri.* "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction." *Speech Communication*, volume 35, no. 3-4:pp. 203–218, October 2001.

[Moreno04a] *Pedro J. Moreno and Purdy P. Ho.* "A new svm approach to speaker identification and verification using probabilistic distance kernels." Technical report, Cambridge Research Laboratory, Hewlett-Packard, January 2004.

[Moreno04b] *Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos.* "A kullback-leibler divergence based kernel for svm classification in multimedia applications." In *Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf* (editors), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[Morik99] *Katharina Morik, Peter Brockhausen, and Thorsten Joachims.* "Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring." In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 268–277. Bled, Slovenia, 1999.

[Mosteller64] *Frederick Mosteller and David L. Wallace. Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, MA, December 1964.

[NIST01] *NIST.* "The NIST year 2001 speaker recognition evaluation plan." `http://www.itl.nist.gov/iad/mig/tests/sre/2001/2001-spkrec-evalplan-v05.9.pdf`, 2001. Electronic document, last accessed: 2013-02-25.

[NIST09a] *NIST.* "NIST speaker recognition evaluation webpage." `http://www.itl.nist.gov/iad/mig/tests/sre/`, 1995–2009. Webpage, last accessed: 2013-02-25.

[NIST09b] *NIST.* "The history of automatic speech recognition evaluations at NIST." `http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html`, May 2009. Webpage, last accessed: 2013-02-25.

[NIST10a] *NIST.* "The NIST year 2010 speaker recognition evaluation plan." `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`, 2010. Electronic document, last accessed: 2013-02-25.

[NIST10b] *NIST*. "Rich transcription evaluation." `http://nist.gov/itl/iad/mig/rt.cfm`, October 2010. Webpage, last accessed: 2013-02-25.

[Pallett02] *David S. Pallett*. "The role of the national institute of standards and technology in DARPA's broadcast news continuous speech recognition research program." *Speech Communication*, volume 37, no. 1–2:pp. 3–14, 2002.

[Pelecanos01] *Jason Pelecanos and Sridha Sridharan*. "Feature warping for robust speaker verification." In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, pp. 213–218. Crete, Greece, June 2001.

[Peskin03] *Barbara Peskin, Jiri Navratil, Joy Abramson, Douglas Jones, David Klusacek, Douglas A. Reynolds, and Bing Xiang*. "Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, volume 4, pp. 792–795. Hong Kong, April 2003.

[Porter01] *Martin F. Porter*. "Snowball: A language for stemming algorithms." http://snowball.tartarus.org/texts/introduction.html, October 2001.

[Reynolds95a] *Douglas A. Reynolds*. "Speaker identification and verification using gaussian mixture speaker models." *Speech Communication*, volume 17, no. 1–2:pp. 91–108, 1995.

[Reynolds95b] *Douglas A. Reynolds, Marc A. Zissman, Thomas F. Quatieri, G.C. O'Leary, and B.A. Carlson*. "The effects of telephone transmission degradations on speaker recognition performance." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 1995 (ICASSP-95)*, volume 1, pp. 329–332. Detroit, USA, May 1995.

[Reynolds00] *Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn*. "Speaker verification using adapted gaussian mixture models." *Digital Signal Processing*, volume 10, no. 1-3:pp. 19–41, January 2000.

[Reynolds03a] *Douglas A. Reynolds*. "Channel robust speaker verification via feature mapping." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, volume 2, pp. II–53–56. Hong Kong, April 2003.

[Reynolds03b] *Douglas A. Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu,*

Jack Godfrey, Doug Jones, and Bing Xiang. "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition." In *Proceedings of the International Conference on Audio, Speech, and Signal Processing*, volume 4, p. 784?787. Hong Kong, 2003.

[Reynolds03c] *Douglas A. Reynolds, Joseph P. Campbell, William M. Campbell, Robert B. Dunn, Terry P. Gleason, Douglas A. Jones, Thomas F. Quatieri, Carl B. Quillen, Douglas E. Sturim, , and Pedro A. Torres-Carrasquillo.* "Beyond cepstra: Exploiting high-level information in speaker recognition." In *Workshop on Multimodal User Authentication*, pp. 223–229. Santa Barbara, California, December 2003.

[Scheffer10] *Nicolas Scheffer and Robbie Vogt.* "On the use of speaker superfactors for speaker recognition." In *IEEE International Conference on Acoustics, Speech and Signal Processing 2010 (ICASSP '10)*, pp. 4410–4413. Dallas, Texas, March 2010.

[Sebastiani02] *Fabrizio Sebastiani.* "Machine learning in automated text categorization." *ACM Computing Surveys*, volume 34, no. 1:pp. 1–47, March 2002.

[Shen09] *Wenzhu Shen, R.P. Yu, F. Seide, and Ji Wu.* "Automatic punctuation generation for speech." In *IEEE Workshop on Automatic Speech Recognition Understanding 2009 (ASRU 2009)*, pp. 586–589. Merano, Italy, December 2009.

[Shinozaki01] *Takahiro Shinozaki and Sadaoki Furui.* "Error analysis using decision trees in spontaneous presentation speech recognition." In *IEEE Workshop on Automatic Speech Recognition and Understanding 2001 (ASRU '01)*, pp. 198–201. IEEE, Madonna di Campiglio, Italy, 2001.

[Shriberg05] *Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Anand Venkataraman, and Andreas Stolcke.* "Modeling prosodic feature sequences for speaker recognition." *Speech Communication*, volume 46, no. 3-4:pp. 455–472, July 2005.

[Shriberg08] *Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Nicolas Scheffer, Andreas Stolcke, and Murat Akbacak.* "Detecting nonnative speech using speaker recognition approaches." In *Proceedings of the Odyssey Speaker and Language Recognition Workshop 2008*. Stellenbosch, South Africa, January 2008.

[Singer04] *Elliot Singer and Douglas A. Reynolds.* "Analysis of multitarget detection for speaker and language recognition." In *Odyssey: The Speaker and Language Recognition Workshop*, pp. 301–308. Toledo, Spain, June 2004.

[Stamatatos01] *E. Stamatatos, N. Fakotakis, and G. Kokkinakis.* "Computer-based authorship attribution without lexical measures." *Computers and the Humanities*, volume 35:pp. 193–214, 2001.

[Stamatatos09] *Efstathios Stamatatos.* "A survey of modern authorship attribution methods." *Journal of the American Society for Information Science and Technology*, volume 60, no. 3:pp. 538–556, 2009.

[Steyvers07] *Mark Steyvers and Thomas L. Griffiths. Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. Laurence Erlbaum, 2007.

[Switchboard] *Switchboard.* "Switchboard: A user's manual." `http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html`. Webpage, last accessed: 2013-02-25.

[Sönmez98] *M. Kemal Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub.* "Modeling dynamic prosodic variation for speaker verification." In *International Conference on Spoken Language Processing (ICSLP98)*, volume 7, pp. 3189–3192. Sydney, Australia, 1998.

[Tritschler99] *Alain Tritschler and Ramesh A. Gopinath.* "Improved speaker segmentation and segments clustering using the bayesian information criterion." In *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pp. 679–682. Budapest, Hungary, September 1999.

[Tsymbal04] *Alexey Tsymbal.* "The problem of concept drift: Definitions and related work." Technical report, Trinity College Dublin, Ireland, 2004.

[Viikki98] *Olli Viikki and Kari Laurila.* "Cepstral domain segmental feature vector normalization for noise robust speech recognition." *Speech Communication*, volume 25, no. 1–3:pp. 133–147, 1998.

[Wang08] *Chong Wang, David M. Blei, and David Heckerman.* "Continuous time dynamic topic models." In *Uncertainty in Artificial Intelligence*, pp. 579–586. Helsinki, Finland, 2008.

[Weber02] *Frederick Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg.* "Using prosodic and lexical information for speaker identification." In *IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 (ICASSP '02)*, volume 1, pp. 141–144. Orlando, USA, 2002.

[Xiang02] *Bing Xiang, Chaudhari Upendra V., Jiri Navratil, Ganesh N. Ramaswamy, and Ramesh A. Gopinath.* "Short-time gaussianization for robust speaker verification." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, volume 1, pp. 681–684. Orlando, USA, 2002.

[Zheng06] *Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang.* "A framework for authorship identification of online messages: Writing-style features and classification techniques." *Journal of the American Society for Information Science and Technology*, volume 57, no. 3:pp. 378–393, 2006.

[Zigel06] *Yaniv Zigel and Moshe Wasserblat.* "How to deal with multiple targets in speaker identification systems?" In *IEEE Speaker and Language Recognition Workshop 2006 (Odyssey 2006)*, pp. 1–7. San Juan, Puerto Rico, 2006.

[Zissman96] *Marc A. Zissman.* "Comparison of four approaches to automatic language identification of telephone speech." *IEEE Transactions on Speech and Audio Processing*, volume 4, no. 1:pp. 31–44, 1996.

[Zliobaite09] *Indre Zliobaite.* "Learning under concept drift: an overview." Technical report, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania, 2009.

List of Figures

2.1 General speaker recognition system setup. 7

2.2 Speaker recognition system setup with background modelling. 8

2.3 Example of a DET-plot comparing 3 systems with different performance. 10

2.4 Levels of speaker recognition cues 14

2.5 MAP-Adaptation for a 4-component GMM 25

2.6 Support Vector Machine example in 2-dimensional space. 26

3.1 Utterance pre-processing for topic based speaker recognition. 36

3.2 Overview for the log-likelihood ratio scoring implementation of the keyword-based speaker recognition system. 38

3.3 Overview for the SVM implementation of the keyword-based speaker recognition system. 40

3.4 Tf-idf feature extraction for SVM keyword-based speaker recognition. 41

3.5 Overview for the topic-model-based speaker recognition system. 44

3.6 Representing documents as topic mixtures. 45

3.7 Updating the token-topic assignments in document x. The black arrow marks the token currently under consideration. 47

3.8 Scoring example with speaker prototype vectors. 51

3.9 Scoring example with nearest neighbour lazy learner. 52

4.1 Histogram of video file durations in the Bundestag corpus. 61

4.2 Histogram of number of running words per speech in the Bundestag corpus. 62

4.3 Example screenshots from the Bundestag videos (taken from Angela Merkel's speech on the 19th of May 2010, 9 a.m.). 63

4.4 Example for the format of the stenographic transcripts (taken from the transcript of the 111th session of the 16th legislative period on the 12th of September 2007, page 11,478). . 64

5.1 Processing chain for the final speaker recognition system that combines voice-, idiolect-, and topic-based recognition. 68
5.2 DET-curves for keyword-based speaker recognition with Log-likelihood Ratio scoring and Support Vector Machine Classifiers. 75
5.3 DET-curves for keyword-based speaker recognition with and without utterance pre-processing. 76
5.4 DET-curves for keyword-based speaker recognition with rank normalisation and tf-idf term weighting. 77
5.5 Equal error rates for topic-model-based speaker recognition for different numbers of topics. 78
5.6 DET-curves for topic-model-based speaker recognition for different distance measures used in scoring. 79
5.7 DET-curves for topic-model-based speaker recognition with different classifiers. 80
5.8 Topic mixtures for the ministers of finance (Wolfgang Schäuble) and education and research (Annette Schavan), with manually labelled topics. 83
5.9 DET-curves comparing the spectral, idiolectal, keyword-based and topic-model-based speaker recognition systems. 84
5.10 Influence of the average test utterance length and extraction position on the equal error rate. 87
5.11 Topic mixtures for an example utterance using all 655 words (left), 20 words form the middle, and the middle 20 words repeated 10 times. 88
5.12 DET-curves for the topic-model and idiolectal systems and their combination. 91
5.13 DET-curves for combining spectral, topic, and idiolectal systems. 92
5.14 Histogram of the number of test files with particular trial error rates. 95
5.15 Histogram of the number of speakers with particular erroneous trial rates. 96
5.16 Influence of speaker interest change on topic-model-based speaker recognition. Models were trained on data from 1999. 98
5.17 Equal error rates for each test set year when models are trained on data from the previous year. 100
5.18 Equal error rates for each test set year when models are trained on data from *all* previous years since 1999. 100

Figures 5.5.4, 5.8, 5.9, 5.12, and 5.13 have been published in [Baum12] and are used according to Elsevier's author's rights.

List of Tables

2.1 Comparison of the different feature types. Training and test material is 2.5 min respectively in all cases. 14

4.1 Evaluation Design for the Bundestag corpus. 60

4.2 Train, test, and development set sizes in the Bundestag corpus. 61

4.3 Number of documents per set. 65

5.1 Example 2×2 contingency table, showing the number of correct/correct, correct/incorrect, incorrect/correct, and incorrect/incorrect trials for systems A and B. 73

5.2 Equal error rates for different distance measures. 80

5.3 Equal error rates for topic-model-based speaker recognition with different classifiers. . . . 81

5.4 Equal error rates for the spectral, idiolectal, keyword-based, and topic-model-based speaker recognition systems. 85

5.5 Contingency table showing the number of correctly and incorrectly classified trials for the topic-model-based and spectral speaker recognition systems. 85

5.6 Topic-based speaker recognition systems' equal error rates according to the number of words per test utterance and position where those words are extracted. 87

5.7 Equal error rates for the individual systems and their combinations. 90

5.8 Relative weights for the fused systems
(read: system A + system B | system A's weight : system B's weight). 93

5.9 Influence of test utterance length as side information in the fusion. 94

5.10 Correlation coefficients for correlation between the erroneous trial count and other attributes of a test file. 95

5.11 Equal error rates for both topic-based systems according to year of test set, trained on data from 1999. 99

5.12 Equal error rates for both topic-based systems according to year of test set, trained on data from the previous year and from all previous years. 101