Methods for Large-Scale Data Analyses of Regional Language Variation Based on Speech Acoustics

Thomas Kisler

München 2021

Methods for Large-Scale Data Analyses of Regional Language Variation Based on Speech Acoustics

Thomas Kisler

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie

der Ludwig–Maximilians–Universität München

vorgelegt von

Thomas Kisler

aus

München

2021

Erstgutachter: PD Dr.-Ing. Florian Schiel

Zweitgutachter: Prof. Dr. Phil Hoole

Tag der mündlichen Prüfung: 08.02.2019

Danksagung

Zuallererst möchte ich mich ganz herzlich bei meinem „BAfH“[1] Florian Schiel bedanken. Durch unzählige Diskussionen und wertvollen Input hat er maßgeblich zum Erfolg dieser Doktorarbeit beigetragen. Außerdem möchte ich mich bei meinem Zweitkorrektor Phil Hoole bedanken, seine Leidenschaft für die Phonetik und die Wissenschaft an sich ist eine große Inspiration. Besonderer Dank geht auch an den Drittprüfer meiner Disputation, Christoph Draxler, durch den ich ans Insitut gekommen bin, vielen Dank.

Besonderer Dank gilt überdies Uwe Reichel für zahllose Treffen und Diskussionen im Verlauf der Dissertation, als Sparringspartner war er für mich unbezahlbar. Weiterhin möchte ich Felicitas Kleber danken, für viel fachlichen Input in Sachen Dialektologie und Phonetik und zudem für zahllose spannende, unterhaltsame und herausfordernde Unterhaltungen während der Mittagspausen.

Ich möchte mich auch herzlich bei allen bedanken, die das Institut über die Jahre zu einem Ort gemacht haben, an dem man gerne gearbeitet, sich aber auch gerne einmal außerhalb der Arbeitszeit aufgehalten hat. Zuerst zu nennen ist hier der Lehrstuhlinhaber Jonathan Harrington. Zudem Klaus Jänsch, für die hervorragende Administration der IPS Infrastrauktur, und Ulrike Vallender-Kalus, der guten Seele des Lehrstuhls. Für Zerstreuung in Pausen und nach der Arbeit möchte ich der gesamten EKN danken - hier besonders Mona Späth, Hanna Jakob, Teresa Schölderle, Katharina Lehner und Elisabet Haas - sowie Manfred Pastätter, Véronique Bukmaier, Raphael Winkelmann, Markus Jochim und Katrin Wolfswinkler. Außerdem möchte ich mich bei allen Korrekturlesern bedanken (alle wurden namentlich schon einmal genannt). Und ebenso natürlich bei all den Anderen, die über kurz oder lang Wegbegleiter am Institut waren. Für die Motivation und den Austausch besonders beim Endspurt möchte ich mich zudem bei Nikola Eger herzlich bedanken.

Für eine wertvolle Sicht von außen möchte ich Peter Dressler danken, der mir in unseren Mentoring-Treffen einen wertvollen Einblick in seine reiche Lebenserfahrung gegeben hat.

Zuguterletzt möchte ich mich ganz herzlich bei meinen Eltern, Sonja und Werner Kisler, und meiner Schwester, Veronika Kisler, für Ihre Unterstützung über die Jahre bedanken.

[1] „A“ steht in diesem Fall für „Advisor“.

Contents

List of Figures

List of Tables

ABI "The Accents of the British Isles"

ACCDIST *Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table*

AISEB "Accent and Identity on the Scottish English Border"

ANN Artificial Neural Network

ARBF ANOVA Radial Basis Function

AS Auditory Spectrum

ASR Automatic Speech Recognition

BAS Bavarian Archive for Speech Signals

CM Confidence Measure

CV Cross Validation

DCT Discrete Cosine Transform

DH *Digital Humanities*

DNN Deep Neural Network

DT Decision Tree

DTW Dynamic Time Warping

ECB East Central Bavarian

EF East Franconian

F0 fundamental frequency

F1 first formant

F2 second formant

FFT Fast Fourier Transform

FPR false positive rate

G2P Grapheme-to-Phoneme

GMM Gaussian Mixture Model

GT *German Today*

HMM Hidden Markov Model

HNR Harmonics-to-Noise Ratio

HTK Hidden Markov Model Toolkit

IDS Institut für deutsche Sprache

IPA International Phonetic Alphabet

IPS Institute of Phonetics and Speech Processing

LDA Linear Discriminant Analysis

LMU Ludwig Maximilian University

LPC Linear Predictive Coding

LRT Likelihood Ratio Test

LSP Line Spectral Pairs

LVCSR Large Vocabulary Continuous Speech Recognition

MAE mean absolute error

MAP Maximum a-posteriori

MAUS the Munich AUtomatic Segmentation System

MCR Mean Crossing Rate

MHS *mitteldeutsch/hochdeutsche Sprachscheide*

MFCC Mel-Frequency Cepstral Coefficient

MF-PLP Mel frequency Perceptual Linear Prediction (PLP)

ML Machine Learning

MOCCA Measure of Confidence for Corpus Analysis

NIR No Information Rate

OvR Overlap Ratio

PD2 PhonDat2

PLP Perceptual Linear Prediction

POS Part of Speech

PRLM phone recognition followed by Language Modelling

ranger Random Forest Generator"

RASTA Relative Spectral

RBF Radial Basis Function

Rfilt RASTA filtering

RF Random Forest

RMSE root mean squared error

ROC receiver operating characteristic

R the R Programming Language

RVG1 Regional Variants of German

SAM-PA Speech Assessment Methods Phonetic Alphabet

SDC Shifted-Delta Cepstral

S&L segmentation and labeling

SMOTER Synthetic Minority Over-sampling Technique for Regression

SMOTE Synthetic Minority Over-sampling Technique

SE Spectral Entropy

STS Semi-Tone Spectrum

SVM Support Vector Machine

SVR Support Vector Regression

SV Spectral Variance

TPR true positive rate

UV Utterance Verification

VC Voicing Candidate

VI Variable Importance

VP Voicing Probability

VU Voicing Final Unclipped

WCB West Central Bavarian

Y-ACCDIST York-*Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table* (ACCDIST)

ZCR Zero Crossing Rate

Chapter 1

Introduction

1.1 General Overview

German dialectology, i.e., the investigation of German varieties (which constitute a regional stratification of linguistic differences), has a long and rich tradition. The systematic research of dialects dates back to the large-scale investigation conducted by Wenker at the end of the 19^{th} century (e.g., Schmidt et al., 2011, p. 85 and Lameli, 2008b, p. 256). Despite the restrictions of the time, Wenker's coverage of the German-speaking area at that time (i.e., the *German Reich* at the end of the 19^{th} century) was an outstanding achievement and his research remains an invaluable resource for the investigation of German dialects. In order to achieve such wide-spread and geographically dense coverage of the German-speaking area, Wenker sent out 2200 questionnaires, to local schools spread over the German-speaking region, of which 1500 were filled out and sent back (Schmidt et al., 2011, p. 101). Teachers were asked to (mostly orthographically) transcribe certain sentences as they would pronounce it using their own respective vernacular or that of their pupils if the teachers did not originate from the respective survey location. However, this indirect gathering of data has been criticized, amongst other things, for the limited character set available for transcription (Schmidt et al., 2011, pp. 71, 98–100, and 109; Barbour et al., 1990, pp. 63–64).

The way in which data is collected for investigating regional variations has changed

since then. It has evolved from an indirect written interrogation of informants (as, e.g., applied by Wenker), to a direct one, in which the interviewer transcribes what is being said while listening to the informant, to a workflow in which informants are recorded during an interview, which is then transcribed based on the recording. This separation of recording and annotation was made possible by technological advancements in the information technology field and, more specifically, recording devices.

The linguistic variation encountered is often visualized using maps. Technological advancements have also changed the way in which maps are generated, as today it is no longer necessary to painstakingly draw maps by hand. So whereas, e.g., Wrede et al. (1927–1956) had to spend vast amounts of time drawing maps manually, nowadays it is possible to use specialized software such as REDE (Schmidt et al., 2008) or Gabmap (Nerbonne et al., 2010) – allowing better visualization and easing potential changes. This, in turn, mad it much easier to create dialectological maps (Goebl, 2010).

The regional difference of a certain linguistic variable is often visualized by a line separating two differing regions on a map. This line is called an *isogloss* (Chambers et al., 1998, p. 89). An example of a well-known isogloss is the Benrath line, which divides Germany into areas where Low German and areas where High German is spoken (e.g., marking the regional position between the two pronunciation variants for the German word <*Apfel*> as [ʔapfəl] or [ʔapəl]).

Many classical dialectological studies, following the tradition of Wrede et al. (1927–1956), rely solely on auditorily-based transcription of a speech signal. Unfortunately, each transcription process, whether it be done while listening to the informant or based on the recording, has two main problems that are hard to overcome. First, a transcriber, no matter how well trained he or she is, will always perceive utterances subjectively and will compensate for small and sudden variations in speech (e.g., compensation for coarticulation, Johnson, 2011, pp. 120–145). This compensation effect is magnified by the fact that field workers are often well-versed in certain dialects, or are even trained dialecticians and, therefore, are inevitably biased. One example of a problem caused by the perception of subjective annotators is an effect called *field worker isoglosses* (Mathussek, 2016). These are apparent changes that occur between two sites. However, these changes do not stem

from a systematic difference in the language, but are instead due to systematic errors in field workers transcriptions (Mathussek, 2016). Second, depending on how detailed and complete the transcription is (phones, words, or whole phrases), it can be an extremely time consuming and, therefore, expensive task.

A field of research that combines (mostly) auditorily based transcriptions and information technology is called *dialectometry*. It not only relies on computational methods for visualization, but also uses them for an automatic grouping of dialects. The term *dialectometry* was coined by Séguy (1973) and dialectometric methods, as stated by Goebl (2010), use numerical methods to evaluate and cartograph regional variation. Goebl (2010) further distinguishes between three different schools of dialectometry: the *Salzburg* school, the *Groningen* school, and the *Athens (USA)* school of dialectometry. All have in common that they use some form of distance metric (e.g., Levenshtein distance as in Nerbonne et al., 2013 or a relative similarity between available attributes as in Bauer, 2004) to characterize the distance between dialects. These metrics are essential when it comes to transforming a given transcript of words to continuous values describing different dialects. An important basic assumption of most dialectometric studies is that linguistic atlases contain all necessary information about geographic distribution of language variation (Goebl, 2010).

The majority of dialectometric studies is, as in traditional dialectology, based on manually created auditory transcriptions. Nevertheless, some studies have additionally incorporated acoustic features, such as formants[1] (Heeringa et al., 2003; Heeringa et al., 2009; Grieve et al., 2013), bark-scaled filter-banks (Heeringa et al., 2003), and the Zero Crossing Rate (ZCR; Heeringa et al., 2009). These acoustic features are insofar preferable as they are objective measures and can be automatically extracted from a speech signal. They are not only reproducible, i.e., extracting these features from the same speech signal several times will always lead to identical feature values, but they are also able to capture tiny differences that listeners may compensate for, e.g., fronted back vowels in fronting contexts (e.g. *nutzen*) are still perceived as back vowels (Harrington et al., 2008). These features are also reproducible even if obtained by an unreliable extraction method (i.e., one that

[1]Energy peaks in a specific frequency band depending on the resonances in the respective vocal tract shape (e.g., through a certain tongue position).

frequently produces errors), e.g., like for formants. A study relating a large set of acoustic features extracted from speech to regional variation has, to my knowledge, not been conducted before.

The extraction of acoustic features, however, is also linked to a given phonetic transcription, which may again be, if created manually, also subjective and time consuming. Fortunately, alignment methods exist that are able to generate phonetic transcriptions based on either an orthographic transcription or a standard pronunciation (cf. Oesch et al., 2017 for an evaluation of several methods). Methods based on orthographic transcriptions have the advantage that such an orthographic transcription, necessary for the alignment, is easier to obtain, as at least for a language like Standard German, a widely accepted orthographic standard exists. It can be assumed that if multiple transcribers work on the same signal, these orthographic transcriptions will feature less variation, even if conducted by untrained individuals, than narrow phonetic transcriptions do. A further advantage of orthographic transcription compared to fine-phonetic transcription is that it is less time consuming (based on experience, an orthographic transcription can be completed 10 to 20 times faster than a phonetic transcription Draxler et al., personal communication, 2016).

In the future, it will be even more important to reduce the time it takes to transcribe a recording or a whole corpus, due to the widespread availability of Information and Communication Technology, including the internet and recording devices. This means that gathering, sharing, and access to speech material will become even easier. Two such projects that use modern technology to collect dialect data by making recordings using smartphones are Scherrer et al. (2012) and Leemann et al. (2018).

Dialect and regional varieties/variants/variation are used as synonyms in this thesis and describe *regional* phonetic and phonological deviations to the standard (as defined in the *Duden Online* 2018), as in contrast to, e.g., sociophonetic variation. This broad and lax definition is sufficient for the studies presented in this thesis and adhering to this definition enables the circumnavigation of trying to define a term for which a short and precise definition has eluded researchers from the start. Barbour et al. (1990, pp. 55–57) state that "In practice, a more pragmatic attitude [to the definition of dialects] tends to be taken, which is that 'a dialect is what the researcher wants it to be'[...]". This describes

the approach taken here. Generally, the definitions agree on the presence of some kind of deviation from the standard correlating with geography. However, they are often not clear on how much deviation, and what kind of deviation, is necessary to constitute a dialect (for more elaborate definitions of dialects, e.g., cf. Barbour et al., 1990, pp. 136–146; Löffler, 2003, pp. 3–9; Schmidt et al., 2011, pp. 53–59).

This thesis proposes a workflow that allows the objective evaluation of regional variation based on predictive models often applied in Machine Learning (ML) using acoustic features extracted from speech signals.

This process consists of a pre-processing step executed during the automatic segmentation and labeling (S&L) that spots erroneous segments and allows the dataset to be limited to presumably correct parts, as well as an evaluation of the importance of acoustic features by training a predictive model based on speech acoustics. The evaluations are based on the *German Today* (GT) corpus (Brinckmann et al., 2008) in Chapters 2 and 3, and on the Kiel (Kohler, 1996; John, 2012) and the PhonDat2 corpus (The ASR Consortium, 1995) in Chapter 4.

1.2 Thesis Contributions and Structure

The contribution of this thesis is five-fold with regards to the above-mentioned steps for pre-processing and evaluating regional language variation in speech. It addresses the following research questions:

- To what extent is automatic S&L suitable for research on regional variation (c.f. Chapter 2)?
- Can the quality of an automatically generated S&L of speech be assessed automatically (c.f. Chapter 4)?
- Is it possible to estimate a speaker's origin based on a short speech sample based on ML methods using only acoustic features (c.f. Chapter 3)?
- Can the acoustic features that enable such an estimation be related to well-known regional characteristics based on output from ML methods (c.f. Chapter 3)?

- Is it possible to efficiently visualize acoustic features which describe regional variation in the geographic space (c.f. Chapter 3)?

In Chapter 2 the effect of manual correction of automatic S&L is evaluated by means of a well-studied dialect phenomenon: the Central Bavarian Lenition. In order to test the effect, an automatic S&L is carried out by the Munich AUtomatic Segmentation System (MAUS), a method regularly applied to data in phonetic studies (recent examples are Stevens et al., 2016; Reubold et al., 2017; Llompart et al., 2017; Montaño et al., 2017). The usual workflow is to automatically align the canonic representation (generated from the orthographic transcription) to the speech signal and afterward manually correct the segments of particular interest. Due to the amount of time manual S&L requires, it is of great interest to see how skipping such a manual correction influences the outcome of the applied metrics.

Chapter 3 builds on the results of Chapter 2 with regards to the validity of the automatic S&L. The experiments are concerned with the classification/regression of the speaker position in the geographic space that is the German-speaking area. Therefore, it is a predictive rather than a descriptive modeling task of regional variation. This kind of estimation could be applied in, e.g., Automatic Speech Recognition (ASR) to improve recognition by an automatic model-selection, one that switches to a model that fits the speaker's pronunciation better. To my knowledge, this study is the first to estimate a speaker origin continuously in the geographic space, as opposed to assigning a categorical dialect label. As the geolocalization is based on speech acoustics, the predictive approach shows if and how much information about a speaker's origin is contained in a speech sample. Moreover, all three experiments in Chapter 3 attempt to connect the features to dialect phenomena reported in the traditional dialectology literature. Additionally, as well as providing estimates on speakers' origins, a detailed analysis is performed on how geographic space is divided by features based on an aggregated, high-information dataset. This part resembles a dialectometric approach, only that in this thesis regional variation is captured solely by the use of speech acoustics, rather than by auditorily-based transcripts

(cf., e.g., Nerbonne et al., 2013) and due to the numeric character of acoustic features, no distance measures have to be applied.

All automatic methods are likely to produce errors. Errors in S&L influence all subsequent steps as a correct phoneme location within the speech signal is paramount for an analysis that is based on acoustic features. However, human labelers, too, make errors when creating orthographic transcriptions and, when transcribing dialects, use words that do not fit the evidence in the signal well. Such an erroneous transcription is likely to lead to a decrease in the quality of the automatic S&L process, due to a mismatch between the phonemes to be aligned vs. the content of the speech signal. In Chapter 4, a method is proposed that incorporates methods and features from the domain of ASR in order to find exactly those segments that have been either wrongly transcribed or misaligned by the automatic S&L process.

Chapter 5 summarizes the main findings of this thesis.

Prior to the experimental chapters, the next section introduces the tools used in all three research chapters.

1.3 Introduction to Relevant Speech Technology

1.3.1 Overview

All three studies rely on two tools: BALLOON, proposed by Reichel (2012), and the Munich AUtomatic Segmentation System (MAUS), described in Schiel (1999). With the help of these tools, developed at the Institute of Phonetics and Speech Processing (IPS) at the Ludwig Maximilian University (LMU) Munich, it is possible to produce an automatic S&L based on a speech recording and its corresponding orthographic transcription. The first tool, BALLOON, performs the Grapheme-to-Phoneme (G2P) conversion and the second tool, MAUS, uses the canonic standard pronunciation form produced by BALLOON and aligns it to the signal. For more convenient access, these two tools are available via web interfaces and web services (Kisler et al., 2017). The whole S&L process is summarized in Fig. 1.1. The way in which MAUS models the signal-text alignment closely resembles

probability-based ASR models. For this reason, a short introduction to ASR will be given before the alignment procedure is outlined.

1.3.2 Automatic Speech Recognition (ASR)

An ASR system tries to decode the true word sequence W that a speaker has uttered, after the speaker has transformed the abstract representation in his or her mind to an acoustic signal by using his or her vocal tract. This acoustic representation, in the form of a speech signal, is presented to the ASR system (Jurafsky et al., 2009, pp. 319–417). From this, certain acoustic features are extracted by the system which are then used to create hypotheses of what the underlying, truly realized utterance W was. From these hypotheses, finally, the most probable hypothesis is selected, by which the system tries to find the word sequence $\hat{W}$, which is identical to the originally uttered word sequence W (Pfister et al., 2008, pp. 327–328).

More formally this means based on the feature vector sequence $X = x_1x_2...x_n$ the ASR system attempts to find the most likely word sequence $\hat{W} = w_1w_2...w_m$ from all existing words from the available vocabulary V:

$$\hat{W} = \underset{W \in V}{\operatorname{argmax}} P(W|X) \tag{1.1}$$

This is called the Maximum a-posteriori (MAP) decision rule. Equation 1.1 can be rewritten using Bayes Theorem[2] to

$$\hat{W} = \underset{W \in V}{\operatorname{argmax}} \frac{P(X|W) \cdot P(W)}{P(X)} \tag{1.2}$$

in which $P(X|W)$ denotes the probability of the acoustic observation X assuming that W is the underlying word sequence, $P(W)$ denotes the probability of the word sequence W according to the language model (meaning an utterance-independent a-priori information about a certain language), and $P(X)$ denotes the probability of the acoustic observation X. As the term $P(X)$ is constant across an utterance W (consisting of i words w) and,

[2] $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

furthermore, can only be approximated, in many ASR systems this term is often dropped without a change in recognition accuracy. This leads to

$$\hat{W} = \underset{W \in V}{\operatorname{argmax}} P(X|W) \cdot P(W) \tag{1.3}$$

By dropping $P(X)$ in Equation 1.3, it does not describe the a-posteriori probability anymore, but rather the joint probability $P(X, W)$ of the acoustic observation X and the word sequence W (Jurafsky et al., 2009, p. 374). Nonetheless, throughout this thesis, the term a-posteriori probability is used, if not stated otherwise.

The acoustic observations X are often available in the form of a 39-dimensional feature space, most often made up of the first 12 Mel-Frequency Cepstral Coefficients (MFCCs), the signal energy, and their Δ (velocity, slope) and $\Delta\Delta$ (acceleration, curvature; e.g., Jurafsky et al., 2009, p. 336).

Using this information it is possible to generate a word lattice, a compact representation of the alternative ASR hypotheses (Woodland et al., 1998). In turn the word lattice is used by the Viterbi decoder (Viterbi, 1967) to estimate the most likely sequence of words $\hat{W}$ in the signal.

1.3.3 Automatic Segmentation and Labeling

Grapheme-to-phoneme conversion with BALLOON

The first step in the alignment procedure is a Grapheme-to-Phoneme (G2P) conversion achieved by using BALLOON (Reichel, 2012). BALLOON is trained using a large pronunciation lexicon and learns how to transform the standard orthographic form of a given language to its phonological, canonical (standard pronunciation) form (cf. the upper left of Fig. 1.1). It learns how to transform differently sized contexts to achieve good transformation accuracy for known or almost known grapheme sequences by C4.5 Decision Trees (DTs) (Quinlan, 1993). When transforming an input sequence, it starts by testing whether the word appears in the lexicon. If it can not be found, it tries to transform the input while iteratively decrementing the phoneme context until a valid conversion can be performed. By using this strategy unseen words can also be converted to a phonological

form, which then is based on the statistical model for the selected language [3]. This strategy is equivalent to n-gram backoff modes in ASR systems. If the conversion is not possible based on a phoneme context, the conversion is performed in single phoneme steps. This ensures that a conversion is always possible.

As just mentioned, BALLOON is based on data from a language-specific pronunciation lexicon. This means that the rules that are used to perform the conversion are data-driven. Two examples of this type of conversion, taken from Kisler et al. (2019), are the conversion of the two sequences <*Abend*> and <*endba*>. Based on the entry in the German lexicon, <*Abend*> will be correctly transformed to /ʔaːbənt/. The non-word sequence <endba>, however, does not have an entry in the German pronunciation lexicon. The conversion is still possible by the aforementioned reduction in context and results in /ʔɛntba/. In this case, the phonological rules of the German language are sufficiently captured by the material in the pronunciation lexicon (one example is the general realization of word-initial vowels with a glottal stop). Therefore, the conversion of this logatome to its canonic form results in a phoneme sequence that a German native speaker would expect.

Segmentation and Labeling with MAUS

The phonological transcript created by BALLOON is used in MAUS to generate a probability graph for all possible pronunciations, based on the information from a training corpus (i.e., all variants that are present in the training data; cf. the upper right of Fig. 1.1). This graph is subsequently enriched with the prior probabilities extracted from the same training corpus (Schiel, 2015; cf. the middle of Fig. 1.1). This phonotactic/phonological model is the equivalent of an ASR systems word lattice.

MAUS's acoustic model P(X) is based on the features signal energy and 12 MFCCs coefficients (C_1 – C_{13}, i.e., leaving out the first coefficient C_0) and their Δ (slope) and $\Delta\Delta$ (curvature) features. The above information is used together by the Viterbi decoder (Viterbi, 1967) that produces the most likely S&L (cf. the bottom of Fig. 1.1). This is

[3] This means that BALLOON also outputs a standard pronunciation form for logatomes that resembles the expected pronounciation of a native speaker of that language – if and only if the model is able to capture all phonological rules.

performed using the Hidden Markov Model Toolkit (HTK) framework (Young et al., 2002; for an exact description of the MAUS technique, cf. Schiel, 1999).

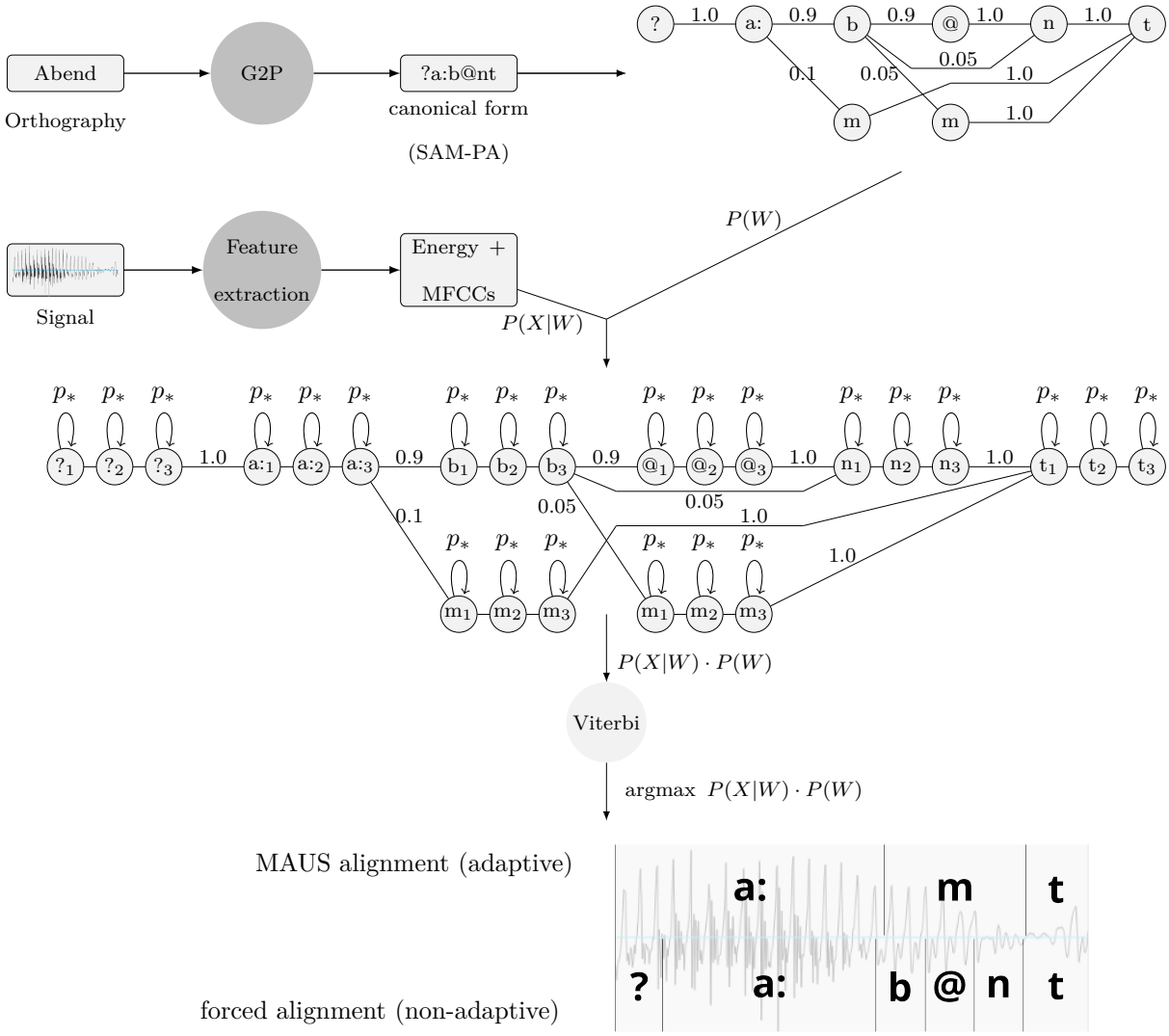


Figure 1.1: Overview of the whole MAUS process aligning the German word "Abend" (evening) to its corresponding speech signal. The two states **start** and **end** are omitted due to space constraints. p_* denotes the respective transition probability between the three states in the HMM that are omitted as well. The forced-alignment (non-adaptive) is an alignment in which the transition probabilities in the first row of the language model are 1.0 (and in all subsequent steps).

Chapter 2

On the Validity of Automatically Segmented Data: an Acoustic Analysis of the Central Bavarian Lenition in the *German Today* Corpus

This chapter is a translation of the following work accepted for publication in German:

Thomas Kisler and Felicitas Kleber (2019). "Zur Validität automatisch segmentierter Daten. Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus". In: *Germanistische Linguistik (Marburg).* Ed. by Sebastian Kürschner, Peter O. Müller, and Mechthild Habermann

2.1 Abstract

The main goal of this study was to evaluate the validity of semi-automatically segmented speech data by analyzing acoustic features primarily related to Central Bavarian lenition in a set of words taken from Bavarian and Austrian speakers' map task recordings taken

from the *German Today (GT)* corpus. A comparison between automatically segmented and manually corrected segment boundaries in a subset of these data shows the same distribution of diatopic and diachronic variation, although the manually corrected data, unsurprisingly, exhibits better separation between, and less variance within, distributions. Our data indicate that potential effects, if anything, tend to be masked rather than exaggerated. Acoustic analyses based on automatically segmented data prove to be a promising, conservative method that promotes and improves the efficient processing of large linguistic corpora.

2.2 Introduction

German dialectology has a long and rich tradition when it comes to creating linguistic atlases of regional variation (e.g., the by now digitally available Wenker-Atlas by Schmidt et al., 2001; Bayerischer Sprachatlas by Hinderling et al., 1996 – 2014, and many more) that show the diversity of Germany's regional varieties. Such dialectological projects are usually long-term ones as they aim to comprehensively map dialects that diverge in small areas on the one hand, whilst, on the other, provide a comprehensive linguistic description of these dialects. To this end, the recording sites have to cover a wide area and be closely-meshed over the dialect regions of interest. The recording and the subsequent preparation of the collected data are time-consuming intermediate steps towards an intended data analysis and the graphical visualization of the data in the form of atlases.

When describing phonological dialect characteristics, auditorily-based transcriptions traditionally play an important role. In this type of transcription, each phoneme is transcribed on the basis of a symbol inventory (e.g., International Phonetic Alphabet, Teuthonista). This method is also time consuming and requires many annotators, not only to process all the given data but, ideally, to have all of it labeled independently by several annotators. This type of multi-person annotation is necessary when attempting to achieve any kind of objective transcription. Every annotator is a person whose perception is subjective, despite their having undergone phonetic training. Hence, it is possible that two people, for example, due to different regional backgrounds, produce differing transcrip-

tions of an identical speech signal (cf. Mathussek, 2016 regarding the problem of field worker isoglosses). In particular, fine-grained phonetic differences, even if they might occur systematically, are not always auditorily perceivable and categorizable. However, in verbal communication (e.g., in word recognition, cf. Hawkins, 2003) these play a linguistically relevant role and are also regarded as a possible cause of diachronic sound change (Beddor, 2009). Analyzing variation on the acoustic level in comparison to analyzing it based on auditory transcription, promises greater objectivity. This is because acoustic parameters are able to capture context-dependent variation for which listeners generally compensate. However, this method generally cannot be used without the carrying out of two pre-processing steps: a) an orthographic transcription, i.e., the representation of the speech signal in standard orthography, and b) a segmentation and labeling (S&L) of the speech signal, i.e., the dissection of the speech signal into single phonetic segments.

The current study has two goals. First, to show that it is possible to study diatopic (and diachronic) variation using a strictly acoustic-based analysis. Second, to present a semi-automatic S&L technique, and to evaluate this method on the basis of the just-mentioned acoustic analysis. The study aims to show the potentials and limitations of this time-saving and more replicable and, therefore, more objective, alternative method. In order to do so, a part of the GT corpus, an already existing, big data collection, was semi-automatically segmented and labeled. The resulting S&L is subsequently used to analyze the phonological dialect feature of Central Bavarian lenition. This well studied dialect feature can be described by the acoustic parameter of duration. Like no other parameter, the duration depends on the segmentation of the signal and, therefore, is ideally suited for the evaluation of an automatic S&L.

The remainder of this paper is organized as follows: Sec. 2.3 first describes the data on which the analysis is based, followed by Sec. 2.4, which introduces the proposed semi-automated pre-processing. Sec. 2.5 describes the dialect feature complementary length in Central Bavaria and its measurable occurrence in the GT corpus in more detail. Sec. 2.6 compares the results of a subset of the data, for which an automatic S&L was obtained, with one that has been manually corrected. Finally, in Sec. 2.7 the pros and cons, as well as the possibilities and limitations of both methods, are discussed.

2.3 The German Today Corpus

The GT corpus serves as the basis for the evaluation of the aforementioned phonological dialect feature. This corpus was compiled as part of the project *Gesprochenes Deutsch* at the Institut für deutsche Sprache (IDS) in Mannheim between 2006 and 2009 (Brinckmann et al., 2008). The goal was a comprehensive survey of diatopic variation in spoken Standard German considering both read and semi-spontaneous speech. The corpus consists of recordings of four speakers at over 160 recording sites, comparatively well-spread across Germany, Austria, and the German-speaking part of Switzerland, as well as a few selected sites in South Tirol, Liechtenstein, East Belgium, and Luxembourg. The speakers of the map task (for an explanation see below) were all local secondary school students (*Gymnasium*), aged between 16 and 20. During the selection of informants, an attempt was made to balance the groups in terms of gender, by recording two male and two female informants at each recording site. The informants had to originate (i.e., have been born and raised) from the region of recording, with the same stipulation applying to at least one of their parents.

The majority of the collected map task data had already been orthographically transcribed by the IDS. This transcription together with the speech signal was fed into the WebMAUS[1] system to create an automatic S&L, which was performed at the Institute of Phonetics and Speech Processing (IPS) at the LMU Munich. At the point of writing, the S&L process had been completed for recordings of 640 informants (328 female, 312 male) from 165 recording sites.[2]

The analyses conducted in the name of the current study are based on the semi-spontaneous map task recordings of 87 speakers from Bavaria and Austria. They can be assigned to the following dialect regions: East Franconian (EF; 23 speakers), West Central Bavarian (WCB; 22 speakers), and East Central Bavarian (ECB; 42 speakers; cf. Wiesinger, 1990 for the separation between West and East Central Bavarian). Fig. 2.1

[1] For a more detailed description please cf. Sec. 2.4.

[2] Due to technical problems during the pre-processing using WebMAUS – such as, for example, broken signal files or erroneous transcriptions – not all speakers who originally took part could be taken into account.

shows the distribution of the recording sites as well as the respective assignment to one of the three dialect regions. In a map task (Anderson et al., 1991), two different speakers enter into a dialog with one another without having visual contact. They are provided with two similar maps that are mostly identical regarding the landmarks[3] they feature. On one of the two maps, a path is drawn from a start to an end point. The task of the speaker in possession of the map containing the path, is to describe the course of the path to the other speaker as exactly as possible. The other participant should draw the path on his or her map (which has no path on it) as accurately as possible, based solely on the other informant's description. The speech material uttered in this setting is semi-spontaneous and consists of multiple realizations, especially with regards to landmarks shown on the map (in the case of the GT corpus, e.g. *Motorrad* – motorcycle, *Metzger* – butcher, *Nüsse* – nuts), which part of the map an informant is referring to, and measurements (e.g. *Ecke* – corner, *Mitte* – middle, *Zentimeter* – centimetre)[4]. Multiple realizations of the same words are especially important for the acoustic analysis so as to be able to distinguish between diatopic and idiosyncratic phonetic variation (for the selection of the target words cf. 2.5).

From this dataset, a subset was extracted that was then manually corrected. This subset contained the same speech material, but only consisted of recordings of 56 speakers (16 East Franconian (EF), 18 West Central Bavarian (WCB), and 22 East Central Bavarian (ECB), cf. Fig. 2.1). A comparison of the automatically created and manually corrected segment boundaries will be performed in Sec. 2.6.2.

[3]Landmarks in the GT corpus are pictures selected based on linguistic considerations, but do not have to possess geographic meaning.

[4]The speech material is not perfectly suited to the evaluation, as it also includes compounds, which are not ideal for comparison. A manual evaluation of stress has shown that this has no influence on the results (for more details, c.f., Sec. 2.6.2). An argument in favor of using this corpus is the fact that it was not specifically designed for the current study, which means that phenomena that can be observed in this general purpose corpus seem to be stable phenomena in spontaneous speech.

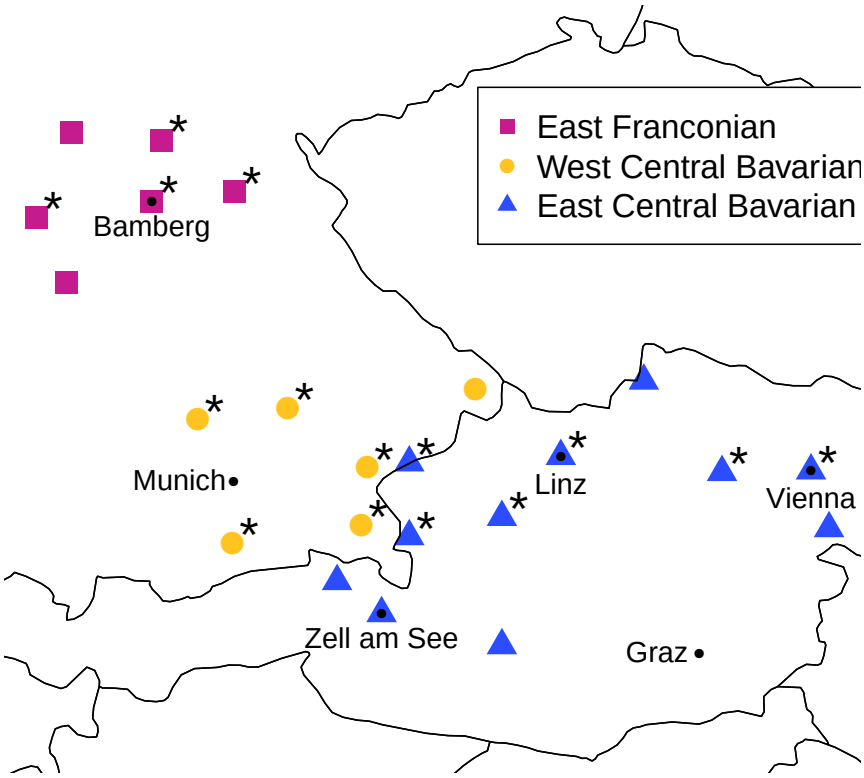


Figure 2.1: Recording sites of the speakers comprising the subcorpus including the assignment to their respective dialect. Locations marked with * indicate that the phonemes' segment boundaries also exist in a manually corrected version.

2.4 Automatic Processing of Speech Signals

Fig. 2.2 presents an overview of the proposed workflow. In the following, all necessary steps that are required to obtain segment boundaries based on a semi-automatic S&L process (in this context also called "labeling" or "annotation") of a particular speech signal[5] are described in more detail. These steps are 1) the manual creation of an orthographic transcription, 2) the automatic S&L using WebMAUS[6], and 3) the acoustic analysis based on the S&L in emuR (Winkelmann et al., 2017). Optionally, it is possible to correct the automatically generated S&L in an intermediate step manually. This kind of correction is described in Sec. 2.6 using the EMU-webApp (Winkelmann et al., 2017). We have omitted this step for the data analysis described in Sec. 2.5.

The term semi-automatic is used, as step 1) the orthographic transcription, usually has to be done manually when dealing with recordings of spontaneous speech. This manual step is necessary, as to our knowledge at the time of writing, no sufficiently good, freely available speech recognition system for German speech exists, especially with regards to regional variants. The subsequent S&L process using WebMAUS, on the other hand, is performed fully automatically (Kisler et al., 2017)[7].

WebMAUS initially performs a Grapheme-to-Phoneme (G2P) conversion using the software tool BALLOON (Reichel, 2012), in which a given orthographic transcription is translated into a SAM-PA[8] transcript of the canonic form (standard pronunciation; cf. Fig. 2.3). By using a digital pronunciation dictionary (here *lexicon*) and trained Decision Trees (DTs)

[5]Here a speech signal might be a recording of several speaking styles as, for example, read speech of word lists or texts, as well as spontaneous speech.

[6]WebMAUS is the name of a web interface, that allows easy access to the two tools BALLOON (grapheme to phoneme conversion) and the Munich AUtomatic Segmentation System (MAUS); both will be explained in greater detail in the following section. The combination of BALLOON and MAUS (together with other tools as, e.g., a syllabification) is also available as a part of the Pipeline service (cf. Kisler et al., 2017).

[7]Therefore, in the context of the datasets based on a S&L created by WebMAUS we will use the term *automatic*.

[8]The Speech Assessment Methods Phonetic Alphabet (SAM-PA) is a machine-readable phonetic alphabet (cf. Wells, 1997).

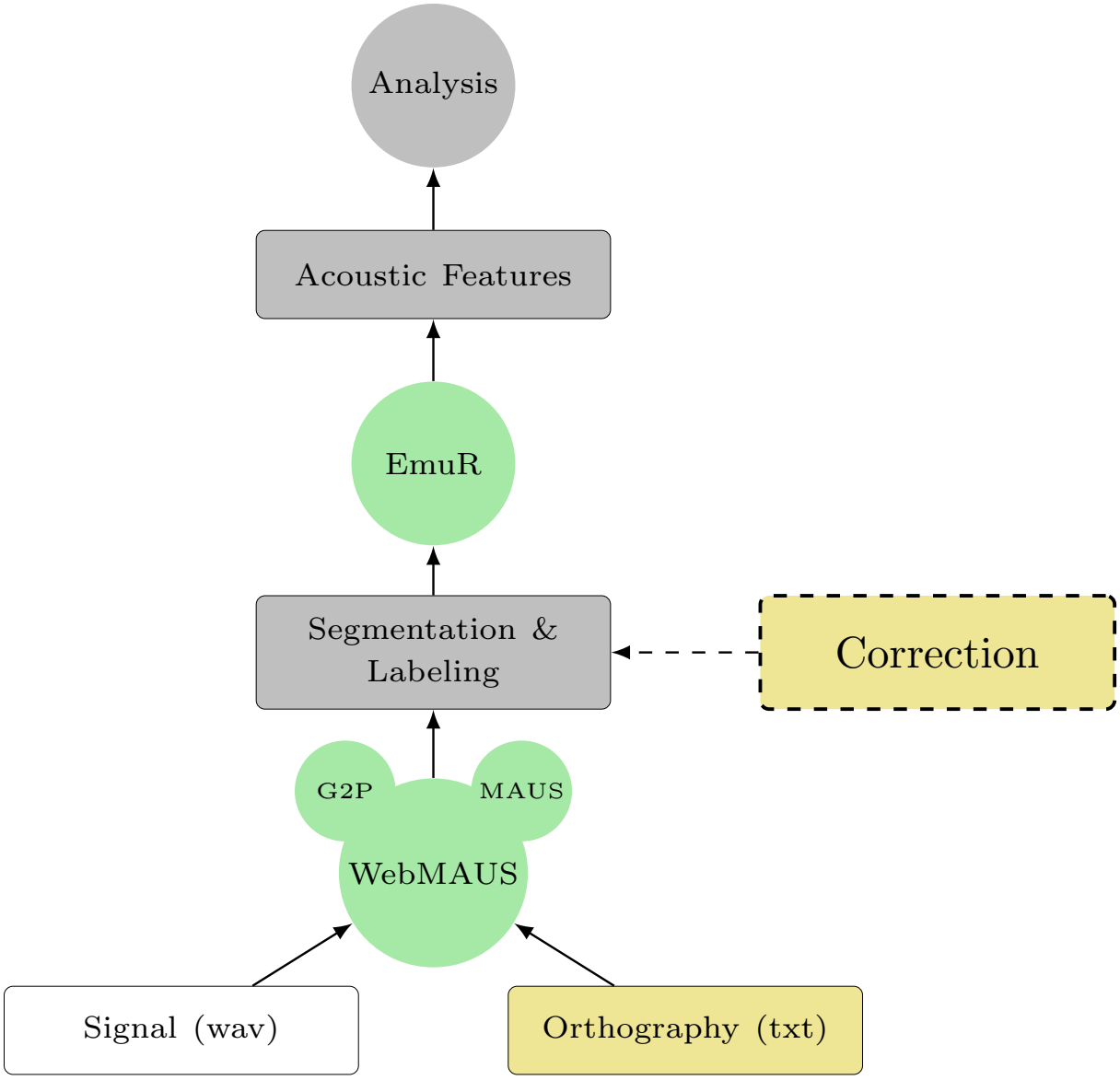


Figure 2.2: Processing steps of the introduced method based on an existing speech signal (white), divided into manual steps (yellow), automatic steps (green), and results of the respective step (gray). The optional step "correction" is outlined in dashes.

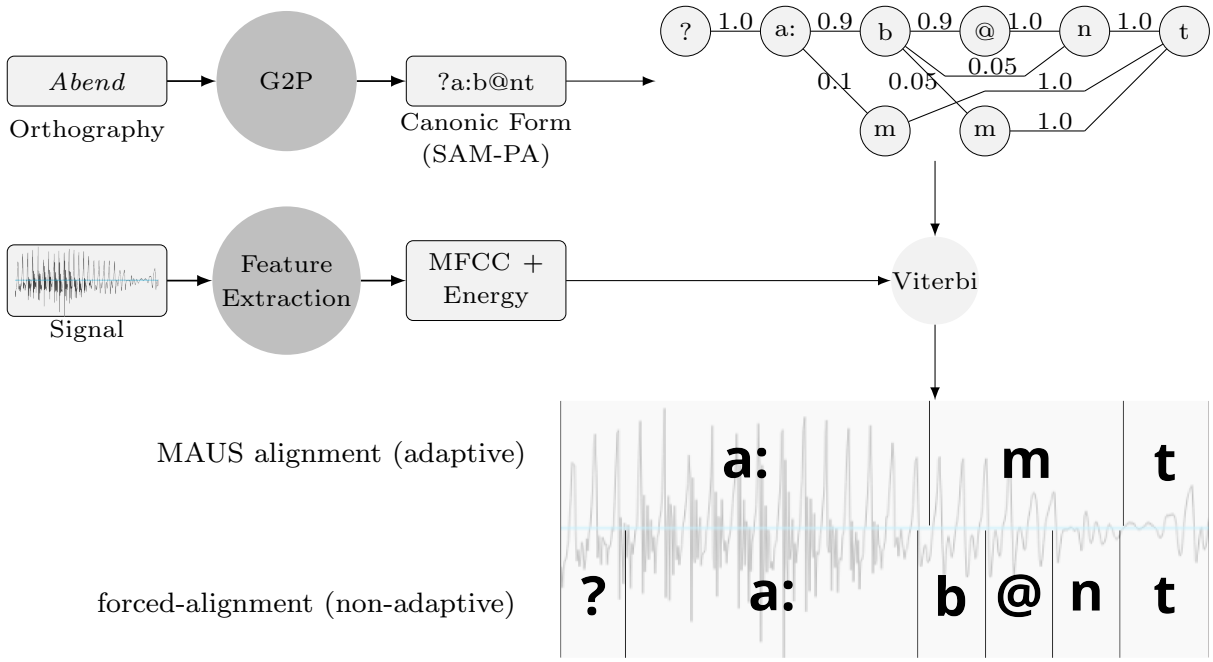


Figure 2.3: Visualization of the workflow of WebMAUS: Grapheme-to-phoneme conversion and extraction of features (upper left), estimation of the most probable phoneme sequence using the Viterbi algorithm (upper right), and the adaptive vs. the non-adaptive symbol-to-signal alignment (bottom).

(Quinlan, 1993) graphemes are converted to phonemes depending on their context. This process starts by using a many-phonemes-covering context, which is successively reduced if necessary (in case conversion is not possible in greater contexts) until a conversion can be performed (Kisler et al., 2017). Therefore, this method also works with words that are unknown to the system (e.g. logatomes). The implemented multi-level process performs an initial check to see if a variant exists in the lexicon that can directly be used to perform the conversion. If the current word is unknown, the system tries to perform the conversion by successively shrinking the grapheme contexts (e.g., <e> in the logatome *klaben* will be translated to /ə/ based on its context). If the conversion cannot be performed using a context (if there are phonemes that have not been observed in the current contexts before) the conversion is performed in single phoneme steps. The rules for the conversion of these types of sequences were extracted from the lexicon during training. Therefore, the rules are data-driven.

Two examples of this process are the conversion of the two sequences of <Abend> and <endba>. <Abend>, an existing German word, will be correctly transformed to /ʔaːbənt/, based on an entry in the lexicon. The non-existing sequence <endba>, however, does not have an entry in the lexicon. By reducing the context of the grapheme sequence, the conversion is still possible and results in /ʔɛntba/. This conversion is possible as the phonological rules of the German language are sufficiently captured by the material in the lexicon (e.g., word-initial vowels are generally realized with a glottal stop).

The transcription generated by BALLOON is then given to MAUS. By using the transcript in combination with rules that have been extracted from a training corpus and its manually segmented data (cf. Kipp et al., 1997), MAUS generates a graph containing all possible alternative pronunciation variants. For the canonic form /ʔaːbənt/ the following pronunciation variants are generated: /ʔaːbənt/, /ʔaːmt/, /ʔaːbmt/, and /ʔaːbnt/. Using the statistical information from training data, MAUS enriches the graph containing the different pronunciation variants by the transition probabilities of the respective phoneme sequence (cf. upper right in Fig. 2.3). This process is carried out as the most likely phone-sequence does not always correspond to the canonic form (which is, however, true in the current example).

In the last step, the lexical-phonological information from the transcript is combined with the acoustic features extracted from the speech signal. MAUS uses the Mel-Frequency Cepstral Coefficients (MFCCs) and signal energy, features that are often applied in speech technology (for more information on the feature extraction in MAUS, cf. Schiel, 1999). By using the Viterbi algorithm (Viterbi, 1967), the posterior probabilities of the most likely pronunciation variant and according segment boundaries can be estimated. In this step, each phoneme is assigned a respective signal section (segmentation) and a SAM-PA symbol that represents it (labeling). Based on the internal modeling of phonemes, the duration of any segment is at least 30 ms and can only be incremented in 10 ms steps (Schiel, 1999).

The just described method is called the *adaptive MAUS alignment*, as the most likely pronunciation variant uttered is aligned to the speech signal. In the case of the example word <Abend> this could be, for example, pronounced as /ʔaːmt/. However, it is possible and sometimes beneficial to prevent MAUS from considering pronunciation variants. This mode is called *forced-alignment*. When using this mode the S&L process is solely based on the canonic form (i.e., in our example: /ʔaːbənt/), regardless of whether the evidence in the speech signal supports a different pronunciation variant or not (cf. the bottom in Fig. 2.3). The adaptive MAUS alignment generally improves the quality of the resulting S&L[9], since in spontaneous speech especially, pronunciation differs from the standard form. These deviations from the norm often occur as reductions and assimilations (e.g., [ham] instead of [haːbən]). In the case of a forced-alignment of the canonic form, each phoneme present in this form needs to be assigned to a part of the signal, in which each phoneme has a minimum duration as mentioned before. Forcing the segmentation of phonemes that are not present in the signal, inevitably leads to a subsequent displacement of all following segment boundaries (and hence a wrong alignment). For the automatic S&L process of the semi-spontaneous data in the GT corpus, we therefore used the adaptive MAUS alignment.[10]

[9] As an example, forced-alignment is more appropriate in laboratory speech where a tendency to hyper-articulation can be observed and realizations closer to canonic forms are more likely.

[10] A comparison with an experimentally executed forced-alignment of the same data resulted in only small differences in the S&L of the target words.

Different S&L output formats can be selected in WebMAUS (e.g., TextGrid, emuDB, etc.). The analyses of the current study are based on the emuDB format, which can be processed by the EMU Speech Database Management System (EMU-SDMS, short EMU; cf. Fig. 2.2). In this format phonemes and phoneme sequences, as well as their hierarchical relationships are explicitly modeled[11] and connected to the segment boundaries.

The EMU system allows complex queries of the S&L information (e.g., the combination of a segmental and suprasegmental level) within a corpus to be performed. Such queries are made available by a component of the EMU system, the emuR software package written for the R programming environment[12] called emuR. In this environment, a multitude of acoustic analyses can be carried out over specified time intervals (i.e., phonemes and phoneme sequences).

2.5 Complementary Length in the Varieties of Central Bavaria

The phonological dialect feature that is used as a showcase in the current analyses is that of Central Bavarian Lenition. In Standard German, both vowel length[13] and consonant strength (also voicing or fortis/lenis contrast) build phonemic oppositions and are freely

[11] Explicit modeling means that an unambiguous mapping of segments of different hierarchical levels is created via links (in contrast to other formats such as Praat's TextGrid, where this type of mapping can only be done implicitly via timestamps). This means that, for example, the syllable [ha] in <haben> is explicitly connected to the phonemes /h/ and /a/. For reasons of data consistency and from an information theoretical point of view, including the fact that this explicit modeling results in the possibility to check the well-formedness of the file format itself, this modeling deserves special attention and recognition.

[12] R is a free programming language that was originally developed for statistical analyses (R Core Team, 2017). Its functionality can be extended by packages that can be installed into the environment. This leads to a vast package infrastructure that enables the user to perform a multitude of tasks within R, including, but not limited, to speech database analyses.

[13] When investigating the phonetic parameter *duration*, we also use the term *vowel length* for phonological opposition. This is especially important as the term *quantity* can, in many cases, be assumed to refer to the primary feature (cf. Becker, 1998; Wiese, 2000).

combinable[14]. Examples of this free combination are *Mieder* /miːdɐ/, *Mieter* /miːtɐ/, *Mitte* /mɪtə/, and *Widder* /vɪdɐ/. In Central Bavarian, not all combinations are possible, as short vowels only occur in front of fortis plosives /p, t, k/ and long vowels only in front of lenis plosives /b, d, g/. Many dialectologists (e.g., Wiesinger, 1990; Scheutz, 1983; Kufner, 1964) regard vowel length to be allophonic as it can be predicted on the basis of the underlying consonant strength of the following obstruent. Bannert (1976), on the other hand, postulates that in Central Bavarian complementary length is a prosodic feature, characterized by a specific vowel-consonant duration ratio.

Phonetic analyses support the model of a duration contrast in consonants in Central Bavarian Dialects (regardless whether complementary length or phonemic consonant length is being referred to, cf. Seiler, 2005), as Bavarian speakers generally lengthen their consonants (i.e., not only plosives but also sonorants) after short vowels (Kleber, 2017). That being said, in Standard German the actual segment duration contributes not only to a distinction between short and long vowels (Ramers, 1988)[15], but also significantly to a fortis/lenis distinction. Lenis plosives have shorter closure phases and are not, or only minimally, aspirated; fortis plosives, in turn, have longer closure phases and the following aspiration additionally contributes to a longer total duration when compared to lenis plosives.

According to Kohler (1979) however, it is not the consonant or closure length alone that captures the Standard German fortis/lenis contrast, but rather a combination of vowel and closure duration, which is called *V/(V+C) ratio*. V corresponds to the vowel length, C to the closure duration of the postvocalic plosive, and V+C to the total duration. The values of V+C are approximately the same regardless of the underlying consonant

[14]In this case they can only theoretically be combined without restrictions. Generally, a tendency can be observed towards a complementary distribution of long vowels and lenis plosives on the one hand and short vowels and fortis plosives on the other – especially in labial and velar plosives (e.g., cf. Kleber et al., 2010).

[15]For the sake of completeness, *syllable cut* (Trubetzkoy, 1939; Vennemann, 1991) should also be mentioned here, in which vowel length is defined according to the coupling between vowel and succeeding consonant. However, this only takes the underlying fortis/lenis opposition into account marginally, which is the reason why we will not further consider this idea in this context.

category (Kohler, 1977). In the case of nasally released plosives especially (e.g., in case of Schwa elision in *mieten* [miːtn̩]) this is the most important acoustic cue. A V/(V+C) ratio of approximately 80% corresponds to a long vowel preceding a lenis plosive and a V/(V+C) ratio of approximately 60% to a short vowel preceeding a fortis plosive (Kohler, 1979, p. 332). This acoustic cue also separates short and long vowels in front of fortis plosives (Braunschweiler, 1997). This cue is not only used in Standard German to produce (and perceive) the phonological opposition, but also in certain Standard German varieties (Saxonian, Central Bavarian; Kleber, 2017).

The goal of the analysis performed on the GT data was to investigate, whether the dialect feature complementary length can be found in data that have been segmented and labeled automatically. A second goal is to investigate whether this dialectal feature is not as frequent in the speech of younger speakers as reported by Moosmüller et al. (2014) for ECB and Kleber (2017) for WCB. This posited latter development might be due to an ongoing sound change influenced by Standard German. In our analysis, the EF speaker group corresponds to a study-internal reference group, as they are supposed to exhibit a different pattern when compared to speakers of ECB and WCB. In this group, we expect, on the one hand, the realization of the vowel length contrast by vowel duration and, on the other hand, a general tendency to lenition, occurring independently of vowel length (Rowley, 1990).

For all three speaker groups (ECB, WCB, and EF as described in Sec. 2.3) we selected the following ten target words from the map task data for analysis for three vowel-consonant combinations:

- *Ecke* /ˈɛkə/, *Mitte* /ˈmɪtə/ – i.e., words with a short vowel (V) in front of a fortis plosive (C:), here and hereafter VC: combinations
- *Nägel* /ˈnɛːgəl/ – i.e., a long vowel (V:) in front of a lenis plosive (C), in the following called V:C combinations
- *Motorrad* /ˈmoːtoːʁˌraːt/ (sometimes also /moˈtoːʁˌraːt/), and *-meter* /meːtɐ/ – i.e., V:C: combinations (only realized as part of compounds[16] in the map task data Zen-

[16]Compounds are not ideal for comparison, because word stress in standard pronunciation does not necessarily need to be on the syllables /me:/ respectively /mo:/ (in contrast to the clearly trochaic words

timeter, Millimeter)

Each of these words was realized at least 200 times in the recordings selected for this analysis. The many utterances of the target words promised a reasonably uniform distribution over the recording sites. Nevertheless, it can be seen in Table 2.1 that for the selected words in the current study, regional differences exist regarding their frequencies in the respective dialect areas.

Table 2.1: Distribution of the analyzed words with regards to the speakers' origin sorted by type of vowel-plosive combination (Comb.).

Comb.	Word	EF	WCB	ECB
V:C:	Millimeter	45	17	146
	Motorrad	63	79	87
	Zentimeter	247	229	423
VC:	Ecke	84	27	173
	Mitte	47	79	132
V:C	Nägel(n)	33	50	123

We calculated the aforementioned V/(V+C) ratio for the 2084 words. It is important to note that in this calculation (and in the remainder of this article) C does not correspond to the duration of the closure phase, as it has in studies conducted by Kohler, 1977 and Kohler, 1979[17], but the total plosive duration. The total duration includes both the closure phase and the aspiration phase. The reason for this was the phoneme-level MAUS alignment, in which the aspiration belongs to the plosive. This combination leads to a generally higher portion of the consonant length in the V/(V+C) ratio in the current study when compared to studies in which C only describes the closure phase.

Ecke, *Mitte*, and *Nägel*, especially in Austrian varieties next to /ˌtsɛntiˈmeːtɐ/ also /ˈtsɛntiˌmeːtɐ/ is possible). However, a manual evaluation of stress has shown that this has no influence on the results (for more details, c.f., Sec. 2.6.2). Additionally, it is possible that a higher syllable count results in a shortening of the vowel length (Klatt, 1973). This can, in turn, shorten the V/(V+C) ratio, although, counter evidence has been presented elsewhere (e.g., cf. Crystal et al., 1990).

[17]In the studies carried out by Kohler (1977) and Kohler (1979), the closure length often corresponded

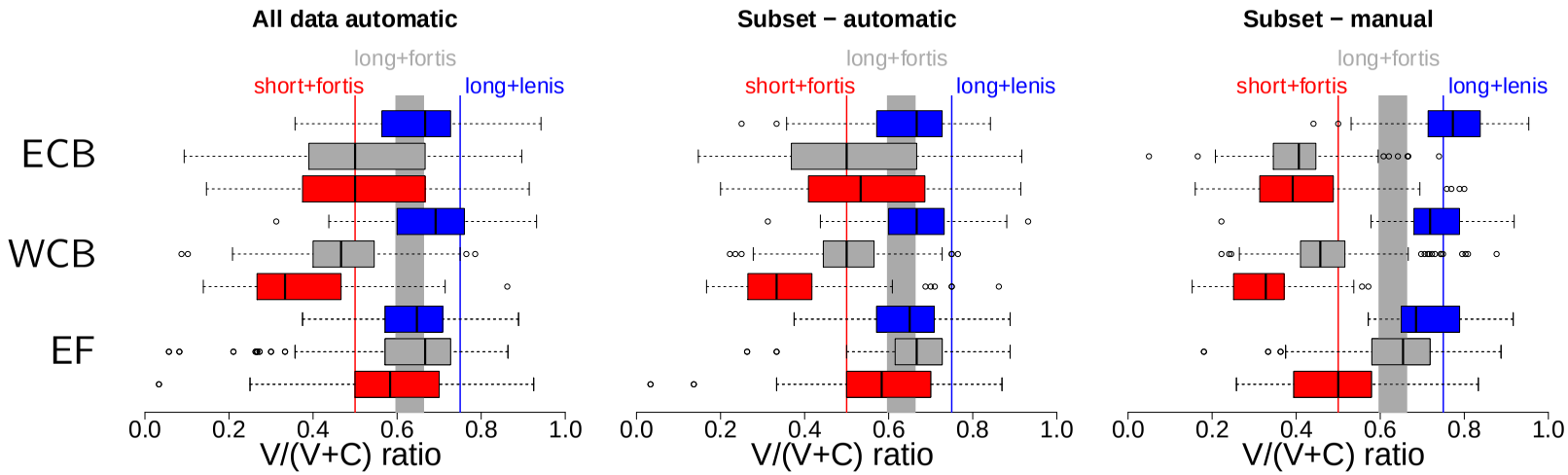


Figure 2.4: V/(V+C) ratio in VC: (red), V:C: (gray), and V:C combinations (blue) separated by speakers originating from EF, WCB, and ECB for the automatic S&L of the complete dataset (left), a subset of the automatic S&L (middle), and the manually corrected data of the subset from the boxplot in the middle (right). The vertical lines correspond to mean values for the V/(V+C) ratio of VC: (short + fortis, red) reported in Braunschweiler (1997), and V:C sequences (long + lenis, blue) reported in Kohler (1979) and Braunschweiler (1997) respectively, the gray bar to the range of values of V:C: sequences reported in Kohler (1979) and Braunschweiler (1997).

Fig. 2.4 shows the V/(V+C) ratio for each of the three speaker groups and the vowel plosive combinations separately. For the analysis described in this section, the leftmost boxplot is of relevance (cf. Section 2.6 for a description of the boxplots in the middle and on the right). This plot shows the V/(V+C) ratio based on the automatic S&L process, for which three observations can be made.

First, a clear separation in at least two V-C length combinations based on the acoustic parameter V/(V+C) ratio can be observed that are in keeping with expectations based on previous studies: the proportional part of the vowel is generally considerably shorter in words with an underlying short vowel when compared to combinations with phonological long vowels in WCB and EF speakers (according to the literature, in ECB speakers these categories collapse, compare for example, Seiler, 2005). This shortening becomes especially apparent in cases in which a lenis plosive follows a long vowel.

Second, the distribution of the V/(V+C) ratios found in the data are shifted to the left (over all speaker groups), when compared to the values found in the literature (cf. the vertical lines in Fig. 2.4; the values for V:C: and V:C are taken from Kohler, 1979, the value for VC: from Braunschweiler, 1997). This indicates that the portion of the vowel in V-C combinations is generally smaller than reported in the literature, which can partly be related to the fact that C corresponds to the total consonant length, not only to vowel length duration. The differing results might stem from the fact that the informants are not speakers of Standard German, but speakers from different dialect regions.

Third, the measured variation of the V/(V+C) ratio of up to 80% of the total duration (which is visualized by the length of the whiskers of the boxplot) is relatively high. This could be evidence of possible measurement errors based on erroneous segment boundaries.

As well as the separation into short and long vowels, a series of other observations can be made based on the left boxplot in Fig. 2.4 that agree with findings in previous studies (which still hold true despite the high level of variation the data exhibits):

- East Central Bavarian (ECB): the V/(V+C) ratio in words with V:C: combinations overlap almost entirely with the ratios in words with VC: combinations. This means

to the total consonant duration, since the plosive was often released nasally and was not aspirated.

that informants from this variety realize both combinations with the same vowel portion. Together with the considerable longer vowel portion in V/(V+C) ratios, this confirms the postulated complementary length for Bavarian speakers holds true for ECB speakers, as long vowels only occur in front of lenis plosives.

Moreover, this data distribution supports models that assume an opposition in phonemic consonant length and allophonic vowel length in Bavarian. Phonetically long fortis plosives, as in words like *Motorrad*, *Zentimeter*, and *Millimeter*, are realized as such, while the Standard German long vowel is produced as a short vowel in ECB.

- West Central Bavarian (WCB): the data gathered from speakers originating from WCB show a clear trend towards a contrast between VC: and V:C: combinations (cf. the significant differences in Table 2.2). Whereas the distribution within ECB speakers is consistent with the description of this dialect feature in the literature, the distribution within WCB speakers indicates a change has taken place when it comes to this feature: long vowels can occur in front of fortis plosives (cf. Moosmüller et al., 2014; Kleber, 2017), as a three-way contrast between VC:, V:C:, and V:C is clearly visible, although the vowel portions are generally smaller compared to standard German. It is interesting that the V/(V+C) ratios in the emerging V:C: category clearly lie in the area in which they would be expected for Standard German pronunciation (as well as in the EF speaker data presented here), which is characteristic for VC: combinations. This resemblance can be interpreted as a relic of the short vowel in front of the fortis plosive. The V/(V+C) ratios based on VC: realizations by WCB speakers, in turn, show shorter proportional vowel duration. These shortened vowel durations cannot be solely explained by the above mentioned general trend towards lower vowel proportion in the V/(V+C) ratio, but are instead a result of the suspected phoneme splitting into long and short vowels (similar to Standard German).

- East Franconian (EF): the smaller short vowel proportions in V/(V+C) ratios compared to the long vowels (which are however not significant, cf. Table 2.2) of the EF speakers indicate a phonemic opposition regarding vowel length. However, the proportional vowel duration does not differ depending on the postvocalic consonant as

it does in Standard German (as well as in the data for ECB and WCB at hand). The strong overlapping V/(V+C) ratios in words with fortis and lenis consonants instead indicate a neutralization of the contrasts in that variety, which is documented in the literature (e.g., cf. Rowley, 1990).

A mixed-model confirms the description of the results (random factors: speakers and word). There are significant main effects for V-C combinations ($\chi^2 = 10.65$; $p < .01$) and region ($\chi^2 = 40.08$; $p < .001$) as well as a significant interaction between the two main factors ($\chi^2 = 53.02$; $p < .001$). Column 1 – 3 in Table 2.2 show the relevant post hoc pairwise comparisons.

The results based on the automatic S&L show that the acoustic parameter of proportional vowel length is an appropriate acoustic feature to use to highlight the existence of phonological opposition (lenis vs. fortis, short vs. long vowel) and the Central Bavarian (also Central German) lenition. This result alone can be taken as evidence for the validity of the automatic S&L. The next section describes an explicit evaluation of this validity.

2.6 Evaluation of the Automatically Segmented and Labeled Data

2.6.1 Comparison of Automatically and Manually Obtained Segment Boundaries

This section will compare the results of the automatic S&L process with a manual correction one. Due to resource constraints, we decided to perform this time-intensive correction on only a subset of the data that was used in the last experiment described in Sec. 2.5. We also decided that this correction should be performed by only one person (instead of multiple people). Using only one annotator to correct the boundaries seemed legitimate, as the goal of this particular study was not to reveal possible annotation differences between multiple human annotators, but between MAUS and manual correction.

However, using only one annotator potentially harbors the problem that the corrected

Table 2.2: Pairwise Bonferroni-corrected post hoc comparisons of different feeatures. "**" denotes a Bonferroni-corrected significance level of $p = 0.01/18$, and "-" non-significant comparisons ($p > 0.05/18$). Correction is abbreviated corr. for space reasons.

Group	Comparison	All data MAUS	Subset before corr.	Subset after corr.
EF	VC: vs. V:C:	**, W=17734	**, W=8112.5	**, W=4536
EF	V:C: vs. V:C	- (p=0.697), W=6098	- (p=0.534), W=3066	- (p=0.015), W=2007
EF	VC: vs. V:C	- (p=0.07), W=1719	- (p=0.08), W=1025.5	**, W=242
WCB	VC: vs. V:C:	**, W=10394	**, W=3763.5	**, W=2518
WCB	V:C: vs. V:C	**, W=1528	**, W=1111.5	**, W=606
WCB	VC: vs. V:C	**, W=439	**, W=319	**, W=73
ECB	VC: vs. V:C:	- (p=0.596), W=97917.5	- (p=0.011), W=34450.5	- (p=0.913), W=30163
ECB	V:C: vs. V:C	**, W=24267.5	**, W=5772.5	**, W=181
ECB	VC: vs. V:C	**, W=11256	*, W=3727.5	**, W=273
VC:	FR vs. WCB	**, W=11633	**, W=7520	**, W=7762
VC:	WCB vs. ECB	**, W=8585	**, W=3149.5	**, W=4561
VC:	ECB vs. FR	**, W=15092.5	- (p=0.047), W=8149	**, W=5502
V:C:	FR vs. WCB	**, W=100409.5	**, W=41288	**, W=39552
V:C:	WCB vs. ECB	**, W=85588.5	- (p=0.868), W=34360.5	**, W=47848.5
V:C:	ECB vs. FR	**, W=68806	**, W=18826	**, W=3799
V:C	FR vs. WCB	- (p=0.131), W=663.5	- (p=0.394), W=470	- (p=0.343), W=462
V:C	WCB vs. ECB	- (p=0.104), W=3575	- (p=0.388), W=1419.5	- (p=0.007), W=883
V:C	ECB vs. FR	- (p=0.932), W=2074.5	- (p=0.935), W=759	- (p=0.005), W=1040

boundaries are biased, and are either too much or too little in agreement with the MAUS segmentation. Nevertheless, the risk that one annotator might produce incorrect boundaries/corrections is accepted in many studies, as boundary correction performed by only one person is common practice in the processing of phonetic speech data. Using a trained phonetician for this task, and one who receives clear instructions (listed in the second to next paragraph) on where to set boundaries in the current case, reduces this risk considerably in our opinion. Therefore, we assume that the boundaries are correctly set in the following and any deviations are negligible from other annotators.

The subset, subject to correction, still contains all the words reported previously, however not from all recording sites. In detail this means that it contains a total of 1265 words from 56 informants from 15 recording sites (cf. recording sites marked with * in Fig. 2.1).

The annotator was instructed to check and, if necessary, correct the segment boundaries (i.e., beginning and end) of the above-mentioned vowels and postvocalic consonants using the EMU-webApp. The important criteria on which the annotator should base decisions were the following: first, the vowel had to start and end with a clearly visible second formant (F2); second, the end of the vowel simultaneously marked the beginning of the plosive, which in turn ended with the visible start of the vocal cord vibrations of the succeeding segment. The annotator also labeled the closure and aspiration phase in all words (to verify the observed left-shift in Sec. 2.5), and labeled the position of the primary stress in all words ending in *-meter* and in *Motorrad* (based on the annotators auditory assessment).

A correction of the segments labels was only considered necessary in five cases and only for realizations of the word *Mitte* (these five cases are distributed over different speakers from all three dialect regions). As the annotator made just a few changes we neither calculated the Inter-labeler Agreement nor Cohens Kappa (Cohen, 1960), both standard measures for the comparison of label differences in multi-person annotations, between MAUS (which counted as one annotator in this case) and the human annotator as both of them would result in high values due to few label differences.

For a comparison of phonetic segmentation, no such widely used measures exist, as they do for the above-mentioned label comparison. Therefore, to compare the magnitude

of the manual correction we used a metric called Overlap Ratio (Paulo et al., 2004). The Overlap Ratio (OvR) specifies the degree to which segments overlap regarding their start and end times and is independent of the length of the segments. Because of this, it seems to be an adequate measure for quantifying the degree of displacement between the manually corrected segment boundaries and the original automatic S&L. The OvR is calculated by:

$$o_r = \frac{t_{ij}}{t_i + t_j - t_{ij}} \tag{2.1}$$

where t_i corresponds to the duration of the phoneme x as segmented by MAUS, t_j the duration of the same phoneme x after manual correction by the annotator, and t_{ij} the duration of overlap between segments t_i and t_j between the MAUS generated segment and the segment after manual correction (cf. Fig. 2.5). The OvR is 1 in case of a perfect overlap (i.e., the automatically and manually obtained segment boundaries are identical)[18] and 0 in case of the absence of any overlap[19].

2.6.2 Comparison Between V/(V+C)-Ratio in Automatically Segmented and Manual Corrected Data

The direct comparison of the V/(V+C) ratios calculated on the reduced dataset, based on the segments from the automatically obtained S&L, and on the full dataset show a similar distribution is shown in Fig. 2.4. This similarity is taken as evidence that the reduced dataset comprises a representative subset (cf. the similar results of the pairwise comparisons in columns 2 and 3 of Table 2.2, which back this hypothesis). All subsequent comparisons refer exclusively to the reduced dataset once before (Fig. 2.4, center plot "Subset – automatic" and Table 2.2, column 4), and once after the manual correction (Fig. 2.4, right plot "Subset – manual" and Table 2.2, column 5).

[18] In the current study an OvR of exactly 1 is accomplished in those cases when the annotator considered a manual correction unnecessary.

[19] Technically the range of the OvR is defined from $[-\infty, 1]$. Values smaller than 0 specify the size of the gap between the segments. However, for the following experiments OvRs < 0 are set to 0, since the size of the gap is of no interest.

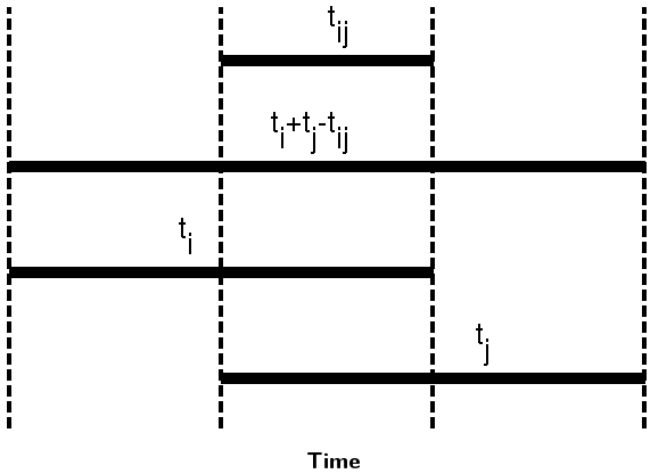


Figure 2.5: OvR of two segments as calculated by Equation 2.1 (Paulo et al., 2004; figure adapted from Kisler et al., 2013a).

The comparison of the dataset before and after manual correction results in similar combination-dependent distributions of the V/(V+C) ratios as reported in Sec. 2.5. The reported results of the pairwise comparisons in column 4 and 5 of Table 2.2 confirm a similar separation in the different linguistic category combinations (long vowel in front of lenis plosive, etc.) by using an acoustic parameter. However, as expected, the separation is clearer and, as seen by the more narrow quartiles in the boxplot in Fig. 2.4, the variation is smaller within the categories in the manually corrected data.

Fig. 2.6 shows the frequency of the OvR per target phoneme between the original and the corrected data. In this histogram, 80% of the phonemes OvRs are over 0.52 (marked by the vertical line). That means 80% of the data have an OvR of more than 0.52. The average of this 80% equates to an OvR of 0.7497. This shows that, in general, the target phoneme was detected (and labeled) correctly by MAUS. However, the exact position of one or both segment boundaries needed a slight correction, even if sometimes only minimally, by the human annotator. In 13 (i.e. 1.06%) cases the OvR had a value of 1, i.e., none of the two boundaries were corrected, and in 69 (i.e. 5.63%) cases only one of the two boundaries was corrected. In 80 of 1227 segments (i.e. 6.52%), the OvR had a value of 0, indicating no overlap between automatic and manual segmentation. In those cases the correct signal

section could not be estimated even roughly. On average, all target phoneme boundaries were shifted 35.40 ms to the left and 25.28 ms to the right. These means encompass all values including outliers.

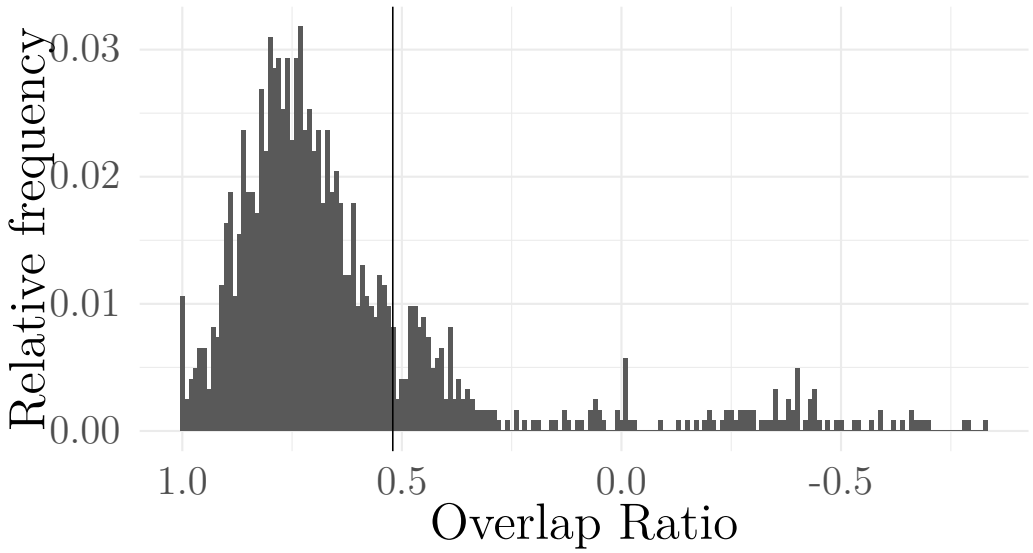


Figure 2.6: Histogram of the OvR between the automatic set segment boundaries and the manual correction. The vertical line marks the position above which 80% of the data lie.

The comparison between the resulting V/(V+C) ratios on the basis of the automatically segmented subset, on the one hand, and the manually corrected subset, on the other, is worth a more detailed analysis. Fig. 2.7 shows the correlation between the V/(V+C) ratios estimated using the automatic segmentation (x-axis) and the V/(V+C) ratios estimated based on the manually corrected segments (y-axis). In cases in which no manual correction was performed (OvR equals 1) the data points lie on the bisecting line. The further away a point is from this line, the bigger the change during the manual correction (and therefore the bigger the resulting difference in the V/(V+C) ratio). The value of the Pearson correlation coefficient of $R = 0.58$ describes a moderate correlation between the V/(V+C) ratios extracted from the two datasets. The achieved correlation means that higher V/(V+C) ratios based on the manually corrected dataset are in general higher for the V/(V+C) ratios estimated for the uncorrected dataset. Moreover, it can be seen that

the difference in the V/(V+C) ratios is not distributed symmetrically around the bisecting line. The scatter plot of the V/(V+C) ratios is skewed towards higher values for the V/(V+C) ratio based on the automatic segmentation.

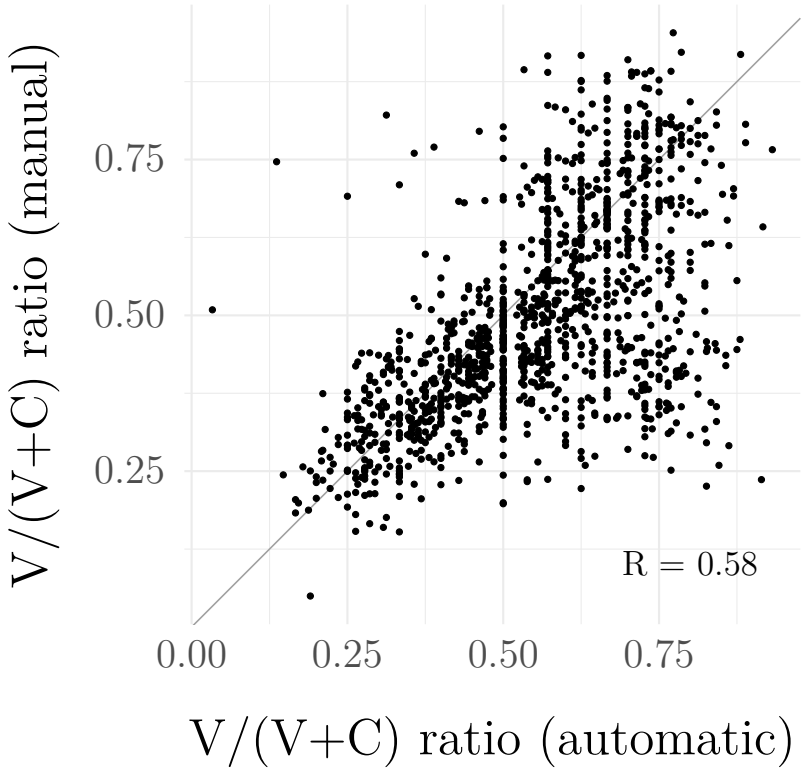


Figure 2.7: Comparison of the V/(V+C) ratios in the automatically segmented and manually corrected data. In the case of perfect overlap, the points lie on the bisecting line. In the lower right, the Pearson correlation coefficient between the V/(V+C) ratio extracted from the automatic segmentation and the manual correction is shown.

The manual post-processing of the data in the EMU-webApp made two more analyses possible that should be mentioned here briefly. First, the fact that the analysis is based on the vowel-plus-total-consonant-length normalized vowel durations (instead of the vowel-plus-closure-phase reported in the literature) showed no significantly different results regarding the distribution of the combination of the linguistic categories. In the vowel-plus-total-consonant-length setting, the vowel proportion was generally smaller for all three V-C combinations. The smaller values are a logical result of the shorter “vowel

plus closure phase" sequence. Second, the manual labeling of stress and the explicit modeling of the hierarchy allowed an investigation of stress patterns, which revealed that the vowels /eː/ in *-meter* and /oː/ *Motorrad* do not often carry the primary stress. For /eː/ only 17% of the analyzed vowels (29 for EF, 9 for WCB, and 92 for ECB speakers) and for /oː/ only 23% of the vowels (exclusively in the data on ECB speakers) carried the primary stress. The fact that the majority of vowels, therefore, carried the secondary stress or were unstressed could have led to shorter vowel durations, as, for example, in words like *Meter* and *Motor*[20]. However, this does not significantly influence the distributions seen in Fig. 2.4.

2.7 Discussion and Conclusion

The present analysis has shown that the automatically segmented semi-spontaneous datasets in the GT corpus are at least suitable for a first inspection of the data. These sets are dialectologically relevant, as it is possible to use them to investigate diatopic variation in spoken Standard German. It can be assumed that recordings capturing regional variation achieve worse automatic segmentation compared to recordings based on Standard German pronunciation as MAUS was trained using data obtained from North German standard speakers from the Kiel Corpus of Spontaneous Speech (Kohler, 1996). However, the current model is able to show regional variation in pronunciation, such as Central Bavarian lenition, if it is present in the data. The present study is the first large-scale semi-automatic acoustic analysis of these lenition phenomena among younger speakers with regional accents common in Bavaria and Austria. The results based on data obtained from WCB speakers also support reports according to which the realization of the combination of long vowels in front of fortis plosives seems to become possible in Central Bavarian dialects (Moosmüller et al., 2014; Kleber, 2017).

The analysis of the lenition phenomenon (both in the speech of Central Bavarian speakers and the EF comparison group) based on the respective phoneme durations, is a feature that directly relies on segment boundaries. Therefore, it is especially well suited for the

[20]This means in words that are not part of a compound.

comparison of the automatically obtained S&L and manual correction. Results based on the automatically obtained duration are influenced by two different factors. On the one hand, the minimal duration of 30 ms slightly stabilizes the results even in cases in which the boundaries are set completely wrong, at least partially. Therefore, an error is limited in that matter, as regardless of what happens to the length of a phoneme, it is never shorter than 30 ms. On the other hand, the segmentation might be too coarse for fine phonetic detail due to the discrete increase in phoneme duration in 10 ms steps. Both restrictions that effect the granularity of the segmentation, result from technical details of the phoneme modeling. However, the region and category dependent V/(V+C) ratios found in all three datasets are taken as evidence for an appropriate validity of the automatic S&L process[21]. The segment boundaries' error resulting in a scattering around zero, based on boundaries that are either erroneously shifted to the left or the right, compensates segmentation errors. This is especially true, as the acoustic content of the segmented signal section does not influence the duration feature.

In contrast to the duration feature, an extraction of formants in a "wrong" vowel (due to erroneous segmentation) could lead to a greater misinterpretation of the data. An example of that would be the interpretation of an open /ɛ/ realization, due to the first formant being extracted from an /a/ that was wrongly segmented and labeled as an /ɛ/.

To verify whether formant data calculated based on an automatic S&L can be used for analysis, the formants of the speakers of ECB and WCB described in Sec. 2.3 were extracted for the words *Mitte* and *Ecke* using emuR. An inspection of this data indicates that the resulting contours of this more sensitive acoustic feature also correspond to phenomena reported in the literature. Fig. 2.8 shows the time-normalized contours of the first formant (F1) based on the automatic S&L in /ɪ/ and /ɛ/ (in Hz). It can be seen that the F1

[21] In manually created and manually corrected segment boundaries a certain variation around real segment position, albeit a smaller one, is also expected. Hence, the positioning of the segment boundaries is per se not exact. Additionally, the positioning heavily depends on the instructions given to the annotator (e.g., the instruction to mark the segment start when the F2 frequency is clearly visible, as done in the current study). They, in turn, are skewed in favor of the phoneme of interest (as e.g., the just mentioned F2 criterion leads to a conservative vowel segmentation, however, potentially to a variable segmentation of the neighboring segments).

values lie in the typical frequency ranges for male and female speakers. Furthermore, F1 is generally lower in *Mitte* than in *Ecke*, since /ɪ/ is produced with a higher tongue position and a more opened jaw. On average the F1 values for ECB speakers are lower than those for WCB speakers. This confirms reports in the literature that short front vowels are produced more tense and therefore with a higher tongue position in Austrian varieties (Cunha et al., 2015).

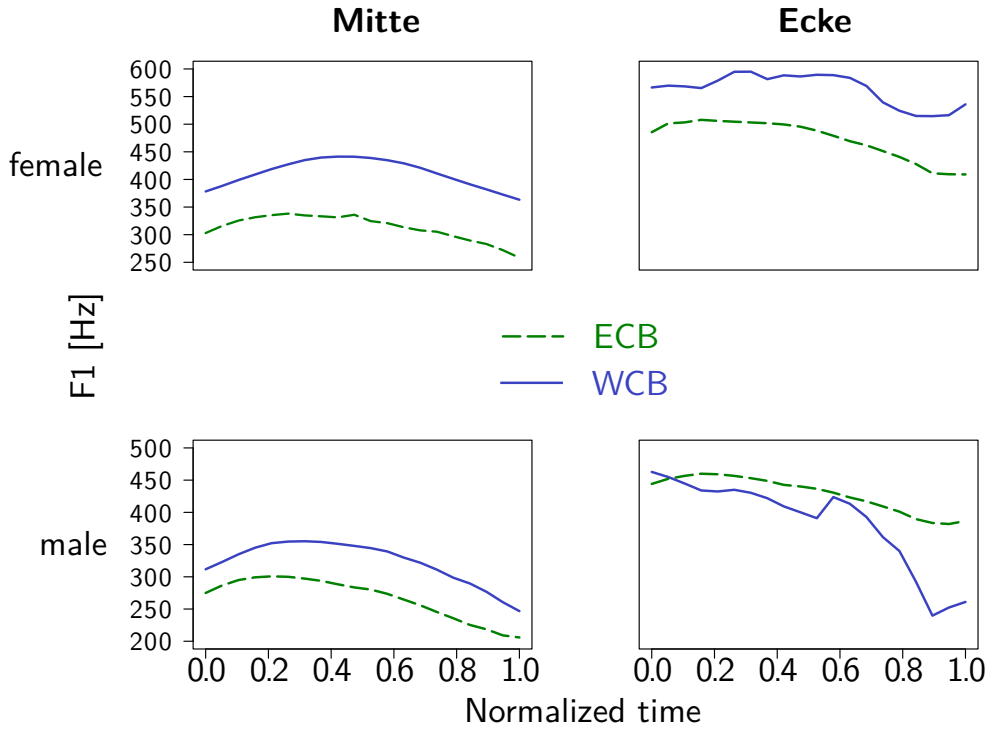


Figure 2.8: Time-normalized F1 contours in the automatic segmented /ɪ/ und /ɛ/ vowels in the words *Mitte* and *Ecke* shown separately for ECB (green; dashed line) and WCB (blue; solid line) for female (top) and male (bottom) speakers.

Analogous to the results of the parameter *duration*, the F1 contours support the validity of the method proposed in this study using acoustic features extracted from the speech signal based on an automatic S&L. However, as already mentioned, more noise is to be expected. This noise is introduced by the extraction of acoustic features in erroneously segmented signal sections. However, once again, the error should produce a scatter around zero (e.g., formant extraction in a wrong signal section will sometimes lead to higher, and

sometimes to lower, values than the ones to be expected).

An automatic S&L of data can be obtained with comparatively little effort, as it is based on the orthographic transcription of a speech signal and not on a fine phonetic one. The resulting broad phonetic transcription allows acoustic analyses to be performed which, in turn allow conclusions to be drawn about differences in fine phonetic detail, for example, about differences between regional varieties. The manual S&L in big corpora is only obtainable by dedicating a lot of resources (time and effort). With the ever-growing size of corpora, this problem will become even more problematic in the future. The automatic S&L can not only be obtained much more quickly, it also has the benefit that the creation of segment boundaries for all segments (not only the ones the researchers are currently interested in) can help answer research questions that arise much later than the creation of the S&L. By the quick and complete S&L of whole corpora (and not only specific target words) a foundation is laid for those future, and at the time of pre-processing, unforeseen studies using the same dataset.

Automatic segmentation errors are without a doubt more frequent and bigger than those for manual segmentation. However, these errors are, in contrast to the errors in manual methods, systematic and objective, as they can be reproduced arbitrarily often, as each execution of the automatic S&L process will produce the exact same boundaries. This objectivity allows for a better comparison of automatically segmented datasets of different research groups, which is more complex for manually created S&L. Further, the acoustic analyses based on automatic S&L are generally more conservative, since the relevant differences are concealed rather than amplified for no reason. These advantages transform the proposed combined method of acoustic analysis based on automatically segmented data into a promising alternative, and this holds true for both linguistic and dialectological questions.

Chapter 3

Geolocalization of Speaker Origins

This chapter builds on the following previously published work:

Thomas Kisler and Florian Schiel (2018b). "Towards a Speaker Localization from Spontaneous Speech: North-South Classification for Speakers of Contemporary German". In: *Elektronische Sprachsignalverarbeitung (ESSV) 2018 - Tagungsband der 29. Konferenz*. Vol. 29. Ulm: TUDpress, pp. 200–207. ISBN: 978-3-95908-128-3

3.1 Abstract

The aim of geographical regression analysis based on phonetic features is to locate a speaker's origin by relating phonetic features derived from a small speech sample to longitude/latitude coordinates. The following chapter contains three experiments to test this type of localization using a large feature set extracted from speech utilizing Random Forests (RFs), Support Vector Regression (SVR), and Decision Trees (DTs).

All experiments use map task data from the "German Today" corpus (Kleiner et al., 2007; Brinckmann et al., 2008), consisting of native German speech. From this semi-spontaneous speech material, an extensive set of 737 features per phoneme is extracted using the openSMILE feature extractor (Eyben et al., 2010). Experiment 1 uses the full feature set, while, based on findings from the first experiment, a reduced subset of 656 features per phoneme is used for experiments 2 and 3.

In a preliminary step (experiment 1), a two-class classification task shows that features extracted from a single /z/ phoneme can be used to correctly assign the origin of 70.37% of the speakers to the North/South half of Germany. This performance can be attributed to the well-known devoicing of phonological /z/ in southern varieties to [z̥] (e.g., Wängler, 1967, p. 143; König, 1989, p. 93–96; Barbour et al., 1990, p. 156). Based on these promising results, a regression analysis was conducted (experiment 2), again using a single uttered phoneme as input. Even for the best-performing phoneme /z/ the improvement over a hypothetical, conservative baseline was only 9.45 km (6.24% of 151.44 km baseline error) for the east-west and 26.69 km (12.65% of 210.89 km baseline error) for the north-south direction.

A third study was conducted to evaluate how suitable the speech material and the extracted features are, and to estimate the lower limit of the mean absolute error (MAE). In this study, data of multiple phonemes occurring in multiple utterances was combined by creating averages from multiple realizations of a single phoneme of each speaker. These averaged vectors were then concatenated for all 33 phonemes that occurred at least once for each speaker. Using a reduced feature set based on highly ranked features according to the measure *Variable Importance (VI)*, an SVR model was able to lower the MAE substantially to 96.14 km in the east-west and to 96.94 km in the north-south direction.

This improvement confirms that the subset contains relevant information in relation to regional variation in the German-speaking area of Central Europe. A binary DT is trained in order to map the subset's features within the geographic space of the German-speaking area and to relate them to already known patterns of regional variation. This mapping of features to a geographic space based on computational methods is somewhat similar to dialectometric approaches (cf., e.g., Nerbonne et al., 2013). However, it differs in the choice of techniques and features. Due to performance issues of the DT regarding the east-west direction, only the north-south direction is analyzed more closely. In this dimension, the division of the geographic space closely resembles traditional dialect boundaries. The top nodes of the DT, again attributed to the devoicing of /z/, are already able to divide the space into a North and South corpus area quite well. The resulting division of the geographic space looks similar to maps based on the variation in linguistic variables in

traditional dialectology. This is taken as evidence for the validity of the approach.

3.2 Introduction

It is commonly assumed that the linguistic features of speakers vary according to where they spent their childhood and where they reside. In German dialectology, it has been shown that the variation of certain linguistic features correlates with geographic distribution (e.g., Wrede et al., 1927–1956; Wiesinger, 1983; König, 1989; Barbour et al., 1990; König, 2005; Brinckmann et al., 2008). The border between two different realizations of a variable (e.g., the different variants of the realization of <Apfel> in one region like [ʔapfəl] and like [ʔapəl] in another) can be visualized as a line on a map indicating the geographic form of the change. This line, which separates two regions by a difference in a linguistic variable, is called an isogloss. The distribution and accumulation of isoglosses allow dialect areas to be defined.

However, dialect areas normally do not possess the clear borders suggested by dialect maps. In many cases, the exact geographic position of a dialect boundary is open for discussion (Wiesinger, 1983; Barbour et al., 1990, p. 85). For instance, moving from one point in Germany to another, one would expect to find places where sudden changes in a single linguistic feature occur, but the overall dialectal variation across the German-speaking area changes semi-continuously along this path as differences of variants accumulate. This assumption is supported, for example, by Haag (1929, p. 19), who states that no two neighboring communities exist that do not recognize differences in their language (even if it is only a different intonation) and Barbour et al. (1990, p. 136), who state that German comprises an enormous amount of varieties, including differences from village to

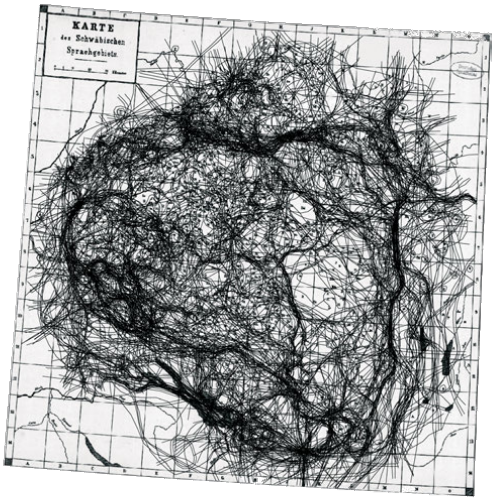


Figure 3.1: Isoglosses from Fischer's atlas (1895) showing the Alemannic-Swabian region (Lameli, 2013, p. 2).

village. Fig. 3.1 shows how many of such linguistic changes can exist in a small area (in this case the Alemannic-Swabian region in 1895) and that those changes can exhibit a semi-continuous character, which stems from the fact that many such single discrete changes occur in close proximity to each other.

In the current study, it is assumed that acoustic features show a regional distribution similar to the linguistic features shown in Fig. 3.1. Hence, it should be possible to exploit these fine-grained changes distributed over a geographic space, to determine the origin of speakers. The method used in the present study relies only on acoustic features extracted from the speech signal and an automatically calculated alignment of phone classes. Except for speaker sex, the system has no access to meta-data or higher-level features (e.g., linguistic or prosodic ones). This restriction is enforced to allow future applications of the proposed method to work fully automatically, with only the speech signal as input. Reducing the amount of subjective data necessary, i.e., avoiding the need for a manual segmentation and labeling (S&L), in which the boundaries are subject to discussion, allows the method to be more objective than other approaches.

It also circumvents the problem of the influence of human experts on the process. An example of such a subjective influence is a field worker isogloss (Mathussek, 2016).

Within the just described paradigm two major questions will be addressed in this study:

1. Is it possible to predict the geographic coordinates (longitude and latitude) of a speaker's origin from a short speech sample, and what would be the average geographic accuracy of such a prediction?

2. Is it possible to automatically divide a large geographic area into regions in which speakers display similar phonetic behavior based on a speech signal sampled from a large and geographically distributed number of speakers?

When answering the second question, another point of interest is whether the resulting division of the geographic space resembles traditional dialect areas and, if so, how these divisions of the underlying features can be visualized effectively.

In the first experiment (cf. Sec. 3.8) a two-class classification was examined. The reason behind taking this intermediate step was to see whether the underlying features

generally carry sufficient information to allow for correct speaker localization, i.e., whether they capture the regional variation sufficiently and which phonemes contribute most to the localization. Two separate two-class classification models were trained to distinguish between speakers from the northern and southern (first model) and the eastern or western (second model) part of the corpus's geographic space. Finally, the experiment aims to identify features that do not have to be taken into account in future experiments.

The second and third experiment deal with (cf. Sec. 3.9 and Sec. 3.10) whether a regression analysis of the speaker origin is possible, how accurate it is, and if it outperforms a theoretical null model. The third experiment (cf. Sec. 3.10) will further try to answer whether it is possible to interpret these features phonetically to explain their existence in the model. In experiment 2 a prediction is made only using data from a single uttered phone. In contrast to this, in experiment 3 the feature vectors of all realizations of a phoneme are averaged, and all averaged vectors of all phonemes uttered by a speaker are concatenated. The two experiments can therefore be viewed as one with sparse information (experiment 2) and one with rich information (experiment 3). Furthermore, the mapping to the geographic space achieved in experiment 3 resembles dialectometrical studies (which also use information technology to map large-scale regional variation to geography automatically).

The outline of the remaining chapter is as follows: the next section gives an overview of the related work, and Sec. 3.4 discusses the approach adopted in this study. Sec. 3.5.1 describes the dataset derived from the corpus *German Today* (GT) provided by the Institut für Deutsche Sprache, Mannheim, Germany, consisting of recordings of contemporary German speech. In Sec. 3.5.2 the pre-processing of the corpus is discussed, Sec. 3.6 explains the extracted features, and Sec. 3.7 gives an overview of the applied Machine Learning (ML) techniques and related metrics. In Secs. 3.8, 3.9, and 3.10 the respective experiments are explained, the results reported and discussed. Finally, in Sec. 3.11 the chapter is summarized.

3.3 Related Work

3.3.1 General Overview

Regression analysis of speaker origins based on speech signals has to my knowledge not been investigated before. This probably has two reasons: a) sufficiently large corpora with well-distributed recording sites have not been available and b) carrying out a regression analysis, as compared to multi-label classification, is more complicated, as the given variation has to be mapped to a continuous scale, not only into a few discrete classes. For the Central European German-speaking countries, problem a) has been solved by the *German Today* corpus becoming available, which includes over 160 recording sites that are well distributed over that area (cf. Sec. 3.5.1).

An area of research that has a similar goal as the research presented here is dialect classification. In this field, the target variable is not speaker position, but a class label. For this, each speaker has to be assigned a predefined dialect class. Most studies in automatic dialect classification are not concerned with the geographic distribution of features, but with the correct assignment of a dialect label (an exception is, e.g., Woehrling et al., 2009). The current study wants to analyze both, to the extent to which the position of a speaker can be estimated automatically, as well as the geographic distribution of the features used to carry out this estimation.

The field of dialect and accent classification using acoustic features from speech signals has been studied extensively and many studies exist, for the English language in particular (e.g., Huckvale, 2004; Huckvale, 2007; Shen et al., 2008; Bahari et al., 2013; Hanani et al., 2013; Brown, 2015). In these studies, a label is assigned to each speaker based on certain acoustic parameters, according to the speaker's accent and the dialect he or she speaks. Research on automatic classification of German-speakers, solely relying on acoustic features, is sparse. The study performed by Stadtschnitzer et al. (2014) is an exception to this. Kisler et al. (2018b) and the work presented here are further contributions.

The different approaches to dialect classification discussed in the following section are categorized into four groups. This grouping results from the combination of speech material

(read or spontaneous) and transcription dependence (transcription needed or not). The following section will discuss the most relevant work to the current approach in each of those four subareas.

A descendant of traditional dialectology called *dialectometry* deals with the efficient and computer-assisted grouping and visualization of variation (Goebl, 2010; Nerbonne et al., 2013). This variation is based mainly on written transcripts, and only a few studies additionally employ acoustic features (such as formants). As the result of the last part of experiment 3 resembles the outcome of dialectometric methods, these methods are briefly outlined in Sec. 3.3.5.

3.3.2 Dialect Classification – Read Speech

Text-Dependent Approaches

Within this approach, the proposed methods extract features based on the transcription of the underlying speech signals. In case the speech material only contains read speech, the transcription is easy to create as it is ideally similar to the text read by the informant.

Huckvale (2004) proposes the *Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table* (ACCDIST) metric, which uses the difference between acoustic parameters between speakers uttering different vowels in British English dialects. In his approach, spectral features are extracted from the first and second half of the available vowel segments describing the spectral envelope. The features are either based on auditory filter banks with 19 filters (Huckvale, 2004) or 20 Mel-Frequency Cepstral Coefficients (MFCC; Huckvale, 2007). For each speaker, these features are used to describe intra-speaker variability, by calculating the distances of the spectral features extracted from the same vowel in different contexts. This method is based on the notion that in British dialects different orthographically identical vowels are pronounced differently in specific words in different regions. An example is the vowel /a/ in “after”, which is pronounced more similar to “a” in “cat” in some regions, in others to “a” in “father”. Speakers are then assigned to one of 14 dialect labels, using the correlations of distances. For the system based on MFCCs, the best accuracy achieved in the “any sex” condition is 86.9%. A

limitation of this method is that the same words have to be read in the same context each time.

A variant of the ACCDIST system, called the York-ACCDIST (Y-ACCDIST) system, was introduced by Brown (2015). She uses the 12 MFCCs along with a Support Vector Machine (SVM) that has been trained using the "Accent and Identity on the Scottish English Border" (AISEB) corpus (Watt et al., 2014). The corpus contains speech from two English and two Scottish varieties located close to the English/Scottish border. The Y-ACCDIST system achieves an accuracy of 86.7% when assigning accents labels. In Brown's study, the Y-ACCDIST system, as was the case for its predecessor the ACCDIST system, the speech material had to be identical for all speakers.

Sinha et al. (2015) use 13 MFCCs, 13 Perceptual Linear Prediction (PLP) coefficients, 13 Mel frequency PLP (MF-PLP) coefficients[1], duration, signal energy, and fundamental frequency (F0) to train an auto-associative neural network. The training was carried out using recordings of read speech by speakers belonging to one of the four main Hindi dialects. Using spectral bottleneck features, they achieved an accuracy of 82% when assigning dialect labels.

Text-Independent Approaches

In contrast to the text dependent approaches, this class of methods does not rely on transcription. Being transcription-independent has the advantage that even if the recorded speech deviates from the text, the system needs no manual input. The methods presented in the following overcome the challenge of there being no transcription either by trying to recognize the correct content of the signal or by modeling the variation over a longer speech sample acoustically.

Hanani et al. (2013) compare different approaches that are based on acoustic and/or phonotactic features, employing phone recognition followed by Language Modelling (PRLM), originally applied to language identification by Zissman (1995). The data consist of read speech from "The Accents of the British Isles" (ABI) corpus (D'Arcy et al., 2004) from

[1]MF-PLP is the combination of PLP and MFCC, in which the PLP coefficients are calculated based on a mel-scaled spectrum.

14 different dialect groups. The lengths of the input signal chunks vary between 30 s and 45 s. Acoustic features were the first 19 MFCCs (including C0) and the Shifted-Delta Cepstral (SDC) coefficients. The phone recognizer necessary for the PRLM approach uses the signal energy plus 12 PLP features, plus their Δ (slope) and $\Delta\Delta$ (curvature) features. To model the different accents acoustically, Gaussian Mixture Models (GMMs) are used. For different combinations, employing either the phonotactic features alone, or in combination with several acoustic features, an accuracy of 89.6% is achieved for the system that fuses all available features. It is worth noting that this performance falls short of the one obtained by an ACCDIST based system, proposed by Huckvale (2004), that was applied by Hanani et al. (2013) to the same data and which achieved an accuracy of 95.18%.

Najafian et al. (2016) propose another text independent system that fuses the features from acoustic accent modeling and phonotactic language modeling using PRLM, to achieve better performance. 19 MFCCs and 49 SDCs are extracted from the speech signal. The phone recognizer uses 12 MFCCs and the signal energy, along with their first and second order derivatives. Here the evaluation is only based on 13 (of the 14 available) accents that exist in the ABI corpus. The input signal lengths vary between 34.5 s and 85.0 s, and the best fused system achieves an accuracy of 84.87%. This performance equals an improvement of roughly 8% compared to the system using only acoustic features (76.76%).

In DeMarco et al. (2013) the authors also use a system based on long read passages from the ABI corpus (presumably using all 14 accents present). They feed a 62-dimensional feature vector, based on SDCs and a warped representation of MFCCs, in an i-vector approach based on Linear Discriminant Analysis (LDA) classification, a nearest neighbors classifier, and SVMs with a cosine kernel. Before the classification takes place, various feature reduction methods are applied (e.g., LDA projection). The fused system achieves an accuracy of 81.05% for 30 s long signal parts.

3.3.3 Dialect Classification – Spontaneous Speech

Text-Dependent Approaches

Methods falling into this category are similar to the text dependent approaches of read speech introduced in Sec. 3.3.2, with the difference that they neither rely on read passages nor on matching content between different speakers.

Woehrling et al. (2009) conducted a dialect identification experiment on French dialect regions. The study used both read (3 minutes) and spontaneous speech (10-15 minutes). Aside from phonetic features such as formant frequencies and voicing, other linguistic features were analyzed such as pronunciation variants (derived from an automatic phoneme alignment), as well as several prosodic features mainly derived from duration measurements and fundamental frequency contours. The best classification rate of 85% was reported for classifying speakers into three major dialect regions using SVMs (82% for five classes). This study is interesting due to three aspects. First, although SVMs yielded the best results, the authors favored a DT for classification due to its interpretability regarding the extracted features. Second, the performance for the tree decreases when more data is added (i.e., more training data does not necessarily yield better results). And thirdly, the performance varies for 3-class vs. 5-class classification over different datasets of read and spontaneous speech, whereas the tasks with fewer classes do not necessarily perform better (which would be expected probability-wise).

In Brown (2015) a variant of the aforementioned Y-ACCDIST system is applied to a 4-way dialect discrimination task based on read and spontaneous speech, again originating from the English/Scottish border. A drop of about one third in classification accuracy for spontaneous (52.5%) compared to read speech (86.7%) is reported. This supports the notion that accent recognition for spontaneous speech is more challenging, than for read speech. A slight increase in accuracy was found when the phoneme context was discarded, probably because the number of observations for each class increased. The features used in the system were the first 12 MFCCs extracted at the vowel midpoint.

The only study on German speech I know of, was conducted by Stadtschnitzer et al. (2014) using the Regional Variants of German (RVG1) corpus (Burger et al., 1998). Using

data from nine large German dialect areas, the authors attempted to predict the speakers' respective dialect membership. They used the speaker identification toolkit ALIZE[2] (Larcher et al., 2013) and found that prediction accuracy was 11.9% when using only acoustic features. Therefore, the achieved performance is just above the level of chance for a 9-class task (11.11%). ALIZE extracts 50 acoustic features (19 MFCCs, 19 MFCC Δ features, 11 MFCC $\Delta\Delta$ features, and Δ Energy). It is interesting to note that the authors also tested a different approach based on a phonemic 4-gram model, which yielded results well above the level of chance for the same dataset (19.2%). This improvement suggests that 'higher' linguistic features (such as phonological, lexical) outperform phonetic features when it comes to dialect classification.

Text-Independent Approaches

Shen et al. (2008) propose a PRLM approach to distinguish between two English and two Mandarin dialects. By fusing the output of the PRLM system and a baseline GMMs system using SDC features, they achieved recognition rates of 81.88% for English and 67.18% for Mandarin using 30 s speech chunks.

Biadsy et al. (2010) propose a system that classifies four different variants of Arabic. Using a sophisticated SVM technique, which exploits the phone context of the material directly in the kernel, the authors reported an average equal error rate of 4.9% in binary classifiers, i.e., four classifiers that discriminate between target dialect or non-target dialect. It is worth noting that, in contrast to Brown, 2015, the authors stress the importance of the phone context. Speech signals first undergo a phone recognition stage that allows the later pairing of data from the same phone type in the discrimination task.

An interesting approach from the field of language identification is Campbell et al. (2006), as this approach employs an SVM using another specialized kernel. To classify 12 different languages, it fuses the results from the SVM and a GMM system, whereas the distributions are modeled using 49 SDC features. For testing, they use 30 s chunks of speech of all 12 languages and achieve an accuracy of 93.6% in the fused system.

[2] All speakers belonging to one dialect class were labeled as the same speaker.

3.3.4 Human Performance in Dialect Classification

When considering automatic dialect classification and geolocalization, it seems important to compare the machine performance to human performance. Therefore, a few studies examining human performance are discussed in the following section.

Draxler et al. (1997) performed perception experiments with two experts[3] and seven non-expert listeners. All listeners were presented with digits from 11 different regions in Germany (mostly according to federal states) from the SpeechDat (M) corpus (*SpeechDat(M): EU-project LRE-63314*). In this setup, the expert listeners achieved a recognition rate of 24.82%, the non-expert listeners 35.55%. The performance of the listeners from different regions varies greatly, however Bavarian non-expert speakers are able to recognize their Bavarian counterpart speakers 100% of the time. This excellent intra-group performance could not be achieved by listeners from other regions. However, the intra-group performance was generally high.

Woehrling et al. (2006) examined the human discrimination performance of six different French regions. 50 participants listened to read and spontaneous speech. They recognized the accent correctly 43.0% of the time, with minor differences between read ($\approx$ 42.3%) and spontaneous speech ($\approx$ 43.7%). Contrary to minor differences in recognition accuracy within different speaking registers, the recognition rate for different dialects again varies greatly.

In Hanani et al. (2013) 24 subjects (aged between 21 and 78) were presented with 20 randomly selected recordings of 14 British accents from the ABI corpus. The length of the chunk of the respective recording presented to the subjects varied between 30 s and 40 s. The subjects achieved 58.24% accuracy, which is significantly poorer than all the automatic systems presented in Hanani et al. (2013).

Human performance in the presented studies is rather poor, and generally worse than in most automatic systems discussed. This poor performance is interesting, as in many domains human are used to set the gold standard (e.g., image understanding, speech recog-

[3]The qualification for expert status was not described in more detail, except that they originated from Bavaria.

nition). This human-made gold standard is then used for modeling systems and often provides an upper limit against which systems have to compete. Setting a gold-standard seems to be more complicated when it comes to regional variation of speech. One surprising aspect of the aforementioned studies is the fact that in Draxler et al. (1997) the listeners with an expert status performed worse than those with non-expert status.

3.3.5 Dialectometry

As mentioned in the introduction (cf. Sec. 1.1), a descendant of traditional dialectology is dialectometry, which applies computational methods to traditional dialectological data. Good overviews of previous research in dialectometry are Goebl (2010), Nerbonne et al. (2013), and Wieling et al. (2015). Dialectometry shows that computational methods – e.g., Levenshtein distance or multi-dimensional scaling – can be successfully employed to relate the resulting groups to traditional dialectological areas (e.g., Gooskens et al., 2004; Goebl, 2010; Nerbonne et al., 2013). For example, the Groningen[4] school of Dialectometry often uses the just mentioned Levenshtein distance to estimate the similarity between variants of the same linguistic variable (e.g., the difference in transcripts between the pronunciation of <afternoon> as [æftənuːn] or [æftərnuːn]).

As mentioned above, most dialectometric studies were concerned with written representations of the speech signal, which were mostly taken from atlases (e.g., Goebl, 2010). In previous studies, only a few acoustic parameters were taken into account (e.g., formants by Heeringa et al., 2009 and timing information by Kisler et al., 2013a). In contrast to many studies based on distance measures, Pickl et al. (2012) moved away from this requirement and, furthermore, postulated a bottom-up view on variation. This view is in line with the work performed in the latter part of experiment 3 of the current study.

[4]For more information on the different schools, cf. (Goebl, 2010). A noteworthy essay describing and criticizing different schools was written by William A. Kretzschmar (2006).

3.4 Chosen Approach

The automatic dialect classification studies mentioned in Secs. 3.3.2 and 3.3.3 are all similar in that they extract features from relatively large chunks of data to perform the classification to achieve good performance. The ACCDIST variants used the realizations of many phonemes in relation to each other, Woehrling et al. (2009) used features derived from a few minutes of speech, and the text-independent systems relied on signal chunks of more than 30 s of speech (that were cut out of longer recordings). In contrast to this, the current study aims to explore whether speaker classification/localization is also possible based on shorter speech samples. Using shorter samples would be beneficial for an application of the findings, e.g., an Automatic Speech Recognition (ASR) system that selects different models based on the speaker's origin. The fact that a model selection improves the recognition in ASR systems has been proved by, for example, by Najafian (2016) for English.

Aside to standard features, like MFCCs or formants, other non-standard features and technology-driven features (those that are usually not used in dialectological/phonetic analyses) are added to the feature set in this study. The advantage of many of these features is that they can be extracted robustly, which is an issue for some widely applied features in dialectological and phonetic studies, such as, for example formants. Additionally, the examined features are restricted to acoustic features, which can be extracted from the speech signal without any other information and only an S&L is required for assigning the feature values to the correct phoneme. Previous studies indicate that higher-level features such as phonotactic or prosodic features improve results. However, using features that can be extracted robustly from a short speech sample might improve applicability, for example, for the aforementioned ASR model selection.

Therefore, in the classification (cf. Sec. 3.8) and regression (cf. Sec. 3.9) analysis only the feature values averaged over the phoneme midpoint plus those 10% to the left and 10% to the right are used. In a later step, the features of multiple phonemes are aggregated and concatenated to form a "dense" feature vector (cf. Sec. 3.10; the density of the vector is comparable to Huckvale, 2007 and Brown, 2015). Taking this approach allow two systems to be compared using sparse information (experiment 2) vs. rich information (experiment

3).

A DT is trained using this "dense" feature vector, and the geographic division resulting from its splitting of the features is related to dialectological phenomena. This is done based on the structure of the features and theoretical knowledge of regional variation. This part of the study tries to replicate what dialectometrical methods achieve, a division of space and a respective visualization of the results. The sole difference between the mentioned dialectometric studies and the current approach is that the current approach only relies on acoustic features. Manual confirmation of the connection between acoustic features (and knowledge about variation in the underlying data) is only carried out for small, random subsamples. Due to the sheer amount of data, it would be unfeasible for the current study to confirm the suspected phenomena across all speakers systematically (e.g., deletion of /ç/ in /ɪç/, pronunciation of word-initial /ç/ as /k/ in words as <Chemie>, etc.).

Kisler et al. (2018b) and experiment 1 in this chapter both examine how accurate speaker origins can be classified into northern and southern parts of the corpus area (cf. Sec. 3.5.1). Both studies use RFs and the same features. The difference between the two lies in the way in which they reach their final decision about which region a speaker originates from. In Kisler et al. (2018b), all realizations of a certain phoneme were taken into account, and the majority vote of those outputs led to the final classification decision. In the current study, this decision is solely based on the RF's output of one produced phoneme, which leads to a less stable prediction. This is because the majority vote in the previous study (Kisler et al., 2018b) also tags a speaker with the label "North" if he or she produces slightly more than 50% of her /z/s voiced, which leads to a wrong decision in many cases in the current study.

3.5 Data and Preprocessing

3.5.1 Corpus

Training and test data were taken from the German speech corpus GT (Kleiner et al., 2007; Brinckmann et al., 2008), which is a valuable resource for the documentation of contem-

porary German. The corpus was recorded in locations distributed over Germany, Austria, Switzerland, and a few sites located in South Tirol, Liechtenstein, East Belgium, and Luxembourg. This area is denoted as the "corpus area" in the following. In each location, four students (mostly two male and two female subjects) from the local secondary school (German: *Gymnasium*) were recorded. All subjects had to have been born and raised in the area, and at least one of their parents had to originate from the region of recording as well. They were aged between 16 and 20 at the time of recording. All recordings were performed using a headset microphone (for more information, c.f. Brinckmann et al., 2008).

In the material analyzed in this study, speakers performed a map task (cf. Anderson et al., 1991) in pairs resulting in semi-spontaneous speech. In map task recordings certain words occur more frequently than others, namely the objects on the map and distance measures. Due to technical issues (e.g., broken signal files and missing transcriptions), not all speakers who took part in the map task contained in the original corpus could be included. Therefore, the current study was carried out on a subset consisting of 641 speakers (328 female, 313 male) from 165 locations (cf. Fig. 3.2). For those speakers, the transcribed speech from the map-task data adds up to $\approx 67\,\mathrm{h}33\,\mathrm{m}$.

Brinckmann et al. (2008) points out that no important traditionally defined dialect area was left out, despite the grid of recording locations not being equally spread over the corpus area. The average distance of the recording sites to their closest neighbor is 41.12 km, whereas the two closest sites are 16.76 km apart and the largest gap is 72.91 km. The distances between sites have a standard deviation of 11.48 km. The GT corpus is the largest available speech corpus with densely spaced recordings over the central European German-speaking regions.

No preselection of phonemes based on words, contexts, or Part of Speech (POS) was performed for the current study. This means that each phoneme class consists of many items realized in various contexts. A preselection would have violated the bottom-up approach. Furthermore, skipping the preselection leads to a higher amount of examples of the available phonemes. which in turn has the chance to improve the models. However, this also means that more noise is contained in the dataset compared to a set only containing carefully preselected data.

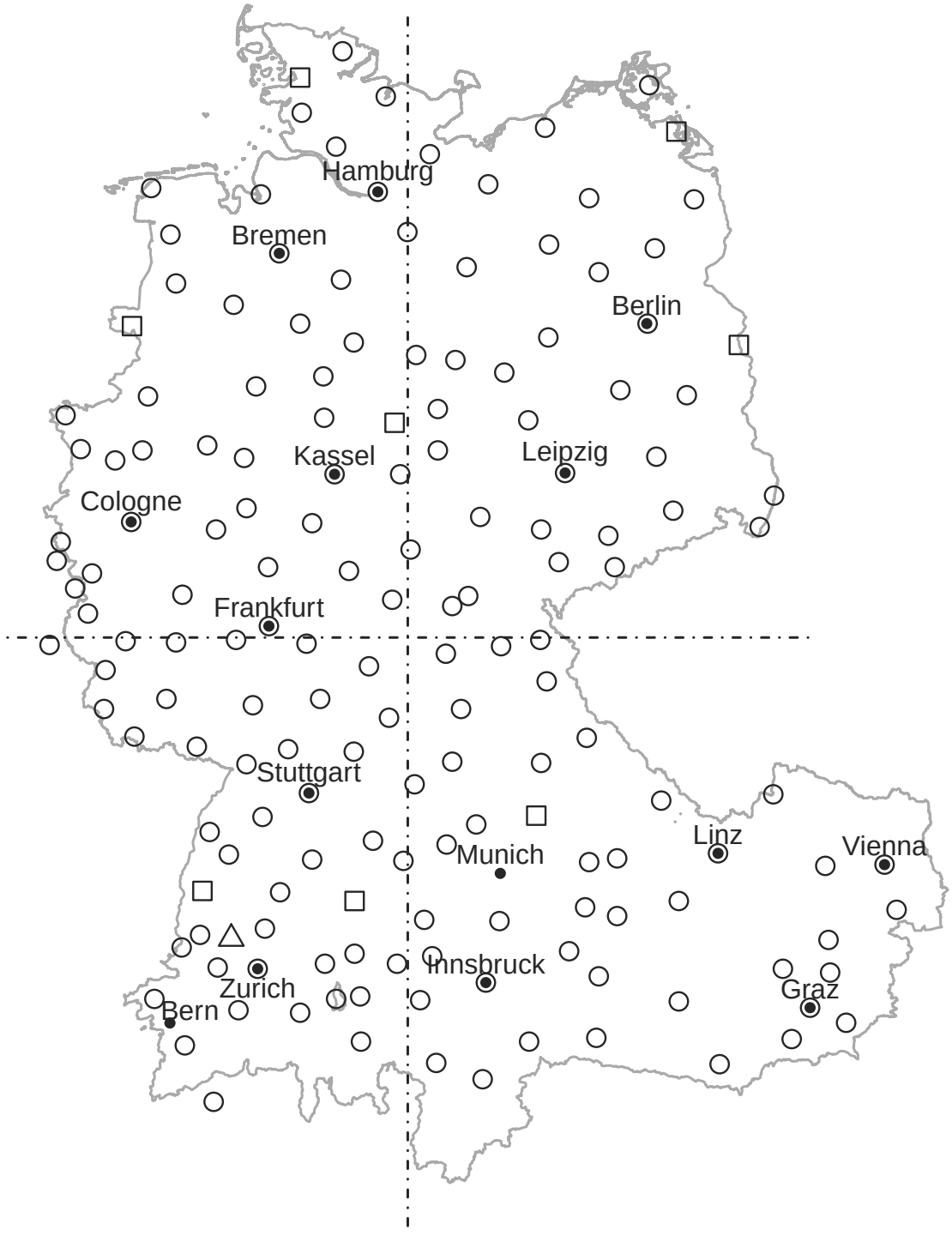


Figure 3.2: Corpus area: 165 recording sites in the GT corpus. At 156 sites four speakers were available for analysis (circle), at eight locations only two speakers (square) and in one location only one speaker (triangle); horizontal line: North and South division; vertical line: East and West division. Black dots indicate reference cities.

3.5.2 Phonetic Segmentation of Speech Material

All recordings in the dataset were transcribed by human annotators. This manually created transcript was then passed to WebMAUS (Kisler et al., 2017) to segment and label recordings into word and phoneme segments (cf. Sec. 1.3.3 for more information on the S&L process). MAUS, the S&L system behind WebMAUS, was applied in the forced-alignment mode to prevent it from changing the canonical transcript. This was done for two reasons: a) the Munich AUtomatic Segmentation System (MAUS) was not trained using dialect data, and therefore might not be suitable to model dialectal variations over the whole corpus area equally well, and b) dialectal differences should be compared using acoustic features alone, i.e., comparing phonetic differences of the same phoneme/phone or evaluating phonological differences based on acoustic features. If MAUS is allowed to change the phone-sequence during alignment, part b) is no longer guaranteed, since, for example, a canonic /z/ might be changed to an /s/ because it is realized without voicing.

Chapter 2 showed that an S&L created by MAUS is a suitable choice when working with regional variation. Further, it showed that applying an automatic S&L means that regional variation is not enhanced. Therefore, it was assumed that no manual correction of boundaries is necessary before feature extraction. Regarding the amount of material available, this would not have been feasible within the limits of this study anyway and would have contradicted the envisaged automatic approach to geolocalization.

The result of applying WebMAUS to the data was a collection of speaker/location labeled segments of 42 different German phonemes. These constitute a subset of the MAUS phoneme inventory for German, which is based on Wells (1997). All phonemes available are listed in Table 3.1 as International Phonetic Alphabet (IPA) symbols.

3.6 Acoustic Features

3.6.1 Overview

In automatic dialect classification, among other linguistic features, several standard acoustic features have been employed frequently, such as

/ə/	/b/	/ɛː/	/ɪ/	/o/	/s/	/v/
/ɐ/	/ç/	/f/	/j/	/oː/	/ʃ/	/w/
/a/	/d/	/g/	/k/	/øː/	/t/	/x/
/aː/	/e/	/h/	/l/	/ɔ/	/u/	/yː/
/aɪ/	/eː/	/i/	/m/	/p/	/uː/	/ʏ/
/aʊ/	/ɛ/	/iː/	/ŋ/	/r/	/ʊ/	/z/

Table 3.1: Available phonemes (42) in the GT corpus represented as IPA symbols.

- MFCCs (e.g., Arslan et al., 1996; Wong et al., 2000; Hansen et al., 2004; Pedersen et al., 2007; Huckvale, 2007; Hanani et al., 2013; Brown, 2015; Sinha et al., 2015)
- signal energy (e.g., Hillenbrand et al., 1993; Arslan et al., 1996; Kat et al., 1999; Hanani et al., 2013; Sinha et al., 2015)
- PLP features (e.g., Hanani et al., 2013; Sinha et al., 2015; Biadsy, 2011)
- formants (e.g., Kat et al., 1999; Huckvale, 2004; Woehrling et al., 2009)
- duration (e.g., Woehrling et al., 2009; Sinha et al., 2015)
- F0 (e.g., Hillenbrand et al., 1993; Kat et al., 1999; Sinha et al., 2015)
- voicing probability (e.g., Woehrling et al., 2009; Finkelstein et al., 2013)

All of these features are part of the feature set employed for geolocalization, including some less used features that can also be extracted from the speech signal. Extracting features that are less frequently used in dialect classification should allow unconventional features to be considered for geolocalization as well. This approach is inspired by the large feature sets used for para-linguistic challenges (e.g., Schuller et al., 2012 or Schuller et al., 2013).

Given that Brown (2015) reported that adding the phoneme context slightly decreased classification performance, no context-dependent features were added.

3.6.2 Overview of Extracted Features

All extracted features applied in the current study are listed in Table 3.2. They were extracted using the openSMILE software package (Eyben et al., 2010) using a Hamming

window (cf. Pfister et al., 2008, p. 65) with a step size of 10 ms and a window size of 20 ms. The openSMILE configuration file that controls the feature extraction can be found in App. A.5.

Mean, RMS, and log energy (3)	Formants (7)
F0[5] (13)	Line Spectral Pairs (LSP) (8)
Auditory Spectrum (AS) (26)	Semi-Tone Spectrum (STS) (96)
AS Relative Spectral filtering (26)	Arbitrary spectral band energies[6] (4)
MFCCs (13)	Spectral roll-off points[7] (4)
Zero Crossing Rate (ZCR) (1)	Spectral centroid and flux (2)
Mean Crossing Rate (MCR)	Spectral Entropy (SE) (1)
Voicing Probabilities (VPs) (6)	Spectral Variance (SV) (1)
(log) Harmonics-to-Noise Ratio (HNR) (2)	Spectral maxpos and minpos (2)
Jitter[8] and shimmer (4)	Spectral slope, skewness, and kurtosis (3)
Chroma features (12)	Spectral harmonicity (1)
Linear Predictive Coding (LPC) (8)	Psychoacoustic sharpness (1)

Table 3.2: The base features that were used in the experiments. Additionally, the short-time functionals slope (Δ) and curvature ($\Delta\Delta$) based on the neighboring frames were used (Eyben et al., 2010). The configuration file to create those features can be found in Appendix A.5. The number of features the description comprises is listed in parentheses behind the name.

The resulting feature vectors were averaged over the 20% midpoint centered region of

[5]The fundamental frequency occurs in four different forms. Raw F0 without thresholding (setting it to 0 if voicing probability τ is smaller than 0.55), F0 after thresholding, smoothed F0, and logarithm of smoothed F0. Additionally, up to four F0 candidates are added (if less than 4 candidates are found, the remaining F0 candidate frequencies are set to 0), the number of candidates found in the current speech sample, and the candidate scores are saved.

[6]The frequency bands are 250 Hz – 649 Hz, 650 Hz – 999 Hz, 1000 Hz – 3999 Hz, and 4000 Hz – 8000 Hz.

[7]The roll-off points are 0.25, 0.50, 0.75, and 0.90, describing the frequency at which k percent of maximal spectral energy can be found.

[8]Three different jitters are calculated: the local frame-to-frame jitter, the differential frame-to-frame jitter, and the envelope of frame-to-frame jitter (for more information cf. Eyben et al., 2010).

each phonetic segment (cf. Fig. 3.3) or the closest vector to the midpoint was selected (in cases in which one feature vector already covered more than 20%).

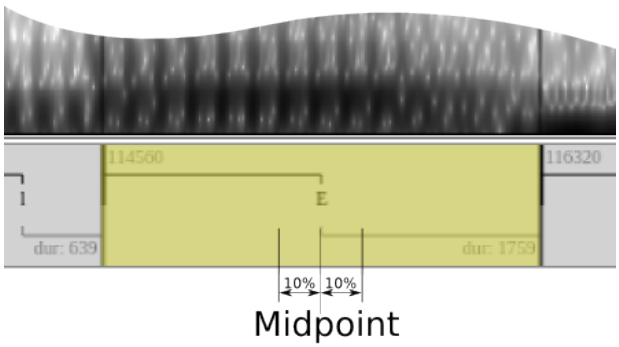


Figure 3.3: The depiction of the extraction point of the features.

Additionally, to the listed features in Table 3.2 the short-time functionals Δ (velocity, slope) and $\Delta\Delta$ (acceleration, curvature) were added as well. The current frame's functional Δ at time t is approximated by (Eyben et al., 2010):

$$\Delta_t = \frac{\sum_{i=1}^{W} i \cdot (v_{t+i} - v_{t-i})}{2 \cdot \sum_{i=1}^{W} i^2} \tag{3.1}$$

where v_{t-i} and v_{t+i} are the feature values of the frames at position $t - i$ and $t + i$ respectively and W specifies half of the window size (in the current study $W = 2$). The $\Delta\Delta$ features where estimated analogously, using the Δ functionals of the two previous and two next frames.

This resulted in a set with $d = 735$ openSMILE features for each phoneme. To this set the phoneme duration and the speaker sex were added. Therefore, the total feature set comprises $d = 737$ features.

The 20% midpoint translates to only a short time segment for most phonemes. For example, for a phoneme with an average duration of 100 ms, on average two feature vectors are combined to form the final vector. This adds up to 30 ms of speech (20 ms Frame 1 + 20 ms Frame 2 – 10 ms overlap). The Δ and $\Delta\Delta$ features cover a larger time period, as they are calculated using the feature values of the surrounding frames. Depending on the duration of the phoneme, Δ and $\Delta\Delta$ feature values sometimes already use information that is located in the transition area between phonemes.

3.6.3 Most Prominent Features

The most important features are described in the following. More information about them can be found in the relevant literature (e.g., Pfister et al., 2008; Jurafsky et al., 2009; Schuller et al., 2014; Eyben et al., 2010).

Duration: Phoneme duration is extracted directly calculated from the S&L using MAUS. For this feature no Δ and $\Delta\Delta$ features are available.

ZCR and MCR: The ZCR describes the rate at which a signal's amplitude changes its sign (from negative to positive and vice versa). The MCR describes, analogously, how often the amplitude crosses the mean. For signals in which the mean is zero, ZCR and MCR are equal. They both describe the periodicity in a speech signal. For both, smaller values are expected in speech segments in which there is voicing (fewer changes). In articulation the zero crossing coincides with the time of equilibrium of air pressure (neither raised nor lowered pressure) around the vocal folds (Gussenhoven, 2004, p. 2).

Voicing Probabilities: An overall of six different voicing probabilities are extracted from the speech signal. For each of the n F0 candidates a voicing probability is estimated based on the subharmonic summation spectrum peak (Eyben et al., 2010, p. 107; in the current study $n = 4$). The feature called Voicing Candidate (VC) in the following is the voicing probability of the best F0 candidate.

Another voicing probability feature is *Voicing Final Unclipped (VU)* of the best F0 candidate. The term *unclipped* is used here to denote that the voicing probability is not set to 0 when it falls below the voicing threshold τ. The "normal" voicing probability is set to 0 for parts in which the voicing probability is below τ (in the current study $\tau = 0.55$).

Spectral Variance and Entropy: Similar to the variance and the entropy of distributions, Spectral Variance (SV) and Spectral Entropy (SE) describe the respective measure of the spectrum (for the calculation of both, cf. Eyben et al., 2010, p. 124 – 125).

Auditory Spectrum (AS): The Auditory Spectrum (AS) is based on the critical band powers of n overlapping mel-scaled triangular filters (in the current study $n = 26$). After calculating the natural logarithm, the equal loudness curve and loudness compression are applied to the spectrum.

For the closely related feature AS Relative Spectral (RASTA) filtering (Rfilt), additionally, RASTA-filtering is applied to the critical band powers (before equal loudness curve and loudness compression is applied). RASTA filtering tries to limit the spectrum to parts containing speech. It does so by filtering changes in spectral band energies that are too slow or too fast to originate from human articulators (for more information, cf. Hermansky et al., 1994).

Semi-Tone Spectrum (STS): As in mel-scaled spectra, in semi-tone spectra the bandwidth increases in higher frequencies. The range of the frequencies covered is 8 octaves, ranging from the first note at 55 Hz to the last note at 14 080 Hz (Eyben et al., 2010). A table that lists the bands can be found in App. A.1.

Mel-Frequency Cepstral Coefficient (MFCC): MFCCs are an efficient and compact representation of a speech signal. They are generated by using the pre-emphasized and windowed[9] frame from which the Fast Fourier Transform (FFT) is calculated. From this spectrum, n bands are extracted using n overlapping mel-scaled triangular filters (in the current case $n = 26$). From each of the mel spectrum values, the logarithm is calculated. Based on this spectral band representation, treating it like a usual discrete time series, a Discrete Cosine Transform (DCT) calculates the decorrelated cepstral coefficients (in the current case, the first 13 coefficients are used, including the first C_0; for more information, e.g., cf. Pfister et al., 2008, p. 296 or Jurafsky et al., 2009, pp. 329 – 336).

Linear Predictive Coding (LPC): LPC uses the notion that subsequent samples in a (speech) signal are not statistically independent. Therefore, the current speech sample $s(n)$ can at least partially be approximated by the past k speech samples $s(n-k), ..., n(-1)$

[9] In this study a hamming window is used.

(in the current study $k = 8$). This is done, for example, to reduce the amount of information that has to be sent over a channel for transmitting speech (e.g., a telephone line). Additionally, LPC allows estimating formant frequencies and bandwidth (Schuller et al., 2014). After LPC analysis only the coefficients and the resulting error term need to be transmitted, whereas the error term can be efficiently compressed (e.g., unvoiced speech parts resemble white noise; for more information cf., e.g., Pfister et al., 2008, p. 81). This leads to a considerable reduction in data size which benefits transmission.

Line Spectral Pairs (LSP): The linear predictor estimated during LPC analysis can be decomposed into a symmetrical and an unsymmetrical part. The LSP coefficients are the zeros of the two resulting polynomials. All zeros lie on the unit circle, the zeros alternate between the symmetrical and unsymmetrical part, and appear in complex symmetrical pairs (hence the name Line Spectral *Pairs*). As all coefficients have the same magnitude, they are more robust against quantization noise than the original LPC coefficients (for more information cf., e.g., Schuller et al., 2014).

3.7 Applied Machine Learning Algorithms and Techniques

3.7.1 Algorithms

Three different widely used non-linear prediction algorithms are employed to combine the advantages inherent to each of them:

- Random Forests (RFs) - predictive power, speed, and feature selection
- Support Vector Regression (SVR) - predictive power
- Decision Trees (DTs) - interpretability of the model

Random Forests

RFs, originally proposed by Breiman (2001), have three major advantages. First, they can be trained quickly, as it is possible to grow the multiple (mostly in the hundreds)

decorrelated DTs in parallel. Second, it is reported that the results of RFs are insensitive to their hyperparameters[10] (Breiman, 2001; Archer et al., 2008; Díaz-Uriarte et al., 2006). And third, they output a feature importance, which allows particularly useful as well as particularly useless features to be spotted (cf. Sec. 3.7.4). Another, albeit smaller, advantage of RFs (this holds true for DTs) is the ability to handle datasets in which the number of dimensions d is much larger than the number of observations n ($d \ggg n$, like in experiment 3 in the current study; James et al., 2014, p. 320).

RFs (like SVMs and DTs) support both classification and regression tasks through minor changes to the algorithm. The first difference is the splitting criterion, which is the *Gini index* for classification, a measure that estimates the impurity in a node (for more information on the Gini index, e.g., cf. James et al., 2014) and the variable response for regression tasks (Wright et al., 2015). The second is the way in which the response for each node individually, and the forest overall (based on single tree predictions), is calculated. In classification tasks, this is done by a majority vote. This means that whatever class appears most often in the selected node is taken as its response. The same is true for the overall prediction of the forest, in which the returned output is the mean of the prediction of individual trees. In regression tasks, the mean of the respective values is used; again for both individual node response and overall forest prediction (Wright et al., 2015).

Another argument for using RFs is that a comparative study on real-world classification problems, revealed that RFs were in many cases the best classifier[11], even outperforming the more complex and slower SVM (Fernández-Delgado et al., 2014). Even though the hyperparameters are reported to be insensitive to changes, the two most important hyperparameters were evaluated in experiments 1 and 2: the number of trees grown in a forest (in the following abbreviated with *ntree*) and the number of features randomly considered at each split (in the following abbreviated with *mtry*). Often dimensionality d is used to define fitting values for mtry. For example $\sqrt{d}$ is the default value for classification and

[10] The term *hyperparameter* refers to the different parameters used to tweak ML algorithms, for example, how they react to misclassifications or how many trees to grow in an RF.

[11] Despite the "No free lunch theorem", which postulates that no algorithm A is better on average on all problems than any other algorithm B (Wolpert, 1996).

$d/3$ for regression tasks for several implementations of the RF algorithm in the R programming environment (R Core Team, 2018). One example is the original implementation of Breiman's algorithm (Liaw et al., 2002a). If the calculation does not result in whole integers, they were rounded to the next smallest integer.

Fully grown trees were generated using the package Random Forest Generator" (ranger) described in Wright et al. (2015) for the R programming environment. The outstanding performance of ranger in comparison to other RF implementations in R is noteworthy (for more information cf. Wright et al., 2015). The default values for minimal node size was left unchanged, which was 1 for classification and 5 for regression.

Support Vector Machines and Support Vector Regression

Support Vector Machines (SVMs) were developed by Vapnik between 1965 and 1995 (Vapnik, 2006) and were originally designed for linear two-class classification. They were extended to non-linear tasks by Boser et al. (1992). For a feature vector with dimensionality p, the SVM tries to find a separating p-dimensional hyperplane, to separate the two classes. If this is not possible, it expands the dimensionality to $p+1$, in which such a separation is always possible. To avoid the computational costs of this high-dimensional feature space, the *kernel trick* is applied (Wang, 2005, p. 24–26). The trick is to compute the inner products not in the feature space (high dimensionality), but by using a kernel function in the input feature space (low dimensionality). The SVM then tries to find the maximum-margin separating hyperplane in the low-dimensional space. Using the notion that examples further away from the hyperplane are less important for its shape, only a subset of training vectors is used to define its shape (called support vectors). Support Vector Regression (SVR; Drucker et al., 1997) is an extension of the SVM algorithm to predict continuous variables. It also uses the kernel trick, analogously to the SVM, to find the best non-linear approximation of the regression function.

In SVM/SVR models that employ a Radial Basis Function (RBF) kernel, two tunable hyperparameters influence the smoothness of the resulting function. These are C, which controls the amount of penalization of deviations from the real values directly in the SVM/SVR model, and γ, which controls the complexity of the projection in the RBF

kernel. The RBF kernel takes the form:

$$k(u, v) = exp(-\gamma \|u - v\|^2) \tag{3.2}$$

For more information on SVM cf. for example Russell et al. (2010) and James et al. (2014), and for SVR cf. for example Schuller et al. (2014).

In the current study, an SVR model is used in experiment 3 to validate the results of a feature selection based on the output of an RF. The SVR was trained using the e1071 package (Meyer et al., 2015) of the R environment, which uses the libsvm library (Chang et al., 2011) internally.

Decision Trees

Decision trees, when compared to the other two algorithms, commonly perform worse (e.g., Woehrling et al., 2009; Fernández-Delgado et al., 2014)[12]. However, they have the benefit of the resulting models in that they are being easy to interpret. In the present study, the trees were trained on features that were ranked highly by the RFs during the training step and were then confirmed to possess predictive power by the SVR. Growing a DT allows for a more straightforward phonetic interpretation of the features compared to RFs and SVR (cf. e.g., Woehrling et al., 2009, p. 2184).

In the current study, a binary decision tree was trained. Binary means that each node results in two alternative paths. The standard splitting criterion applied in the used R package rpart is the Gini index (for more information on rpart cf. Therneau et al., 2018). The final prediction of the tree in regression tasks, like in RFs, is calculated by the mean of the instances residing in the final node.

[12]Hastie (2014) states that the algorithms usually perform in the following order: Boosting $\succ$ Random Forests $\succ$ Bagging $\succ$ Single Tree.

3.7.2 Performance Metrics

Metrics for Classification

For the classification problem, the performance metrics reported are *accuracy*, *precision*, and *recall*. They are defined as

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{3.3}$$

$$\text{Precision} = \frac{t_p}{t_p + f_p}, \tag{3.4}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \tag{3.5}$$

where t_p denotes the true positives, f_p the false positives, t_n the true negatives, and f_n the false negatives.

Precision and recall are necessary as accuracy alone is not a good measure of the classifier quality. This is true for cases in which the classifier output is skewed towards predicting one class more often than another. This happens in cases in which the amount of samples for the two-classes is skewed towards one class and the model only learns to predict one class (i.e, overfits the data). In these types of cases, the accuracy will yield good results despite only outputting one class due to the skewed class distributions. However, this is probably not desired for most applications.

An example of a skewed distribution would be 9.900 observations for class A and 100 observations for class B. If a classifier always outputs class A, the classifier would achieve a high accuracy of 0.99 (out of 1). In this example, 0.99 is called the No Information Rate (NIR). It is defined as:

$$\text{NIR} = \frac{c_{maj}}{t_p + t_n} \tag{3.6}$$

where c_{maj} is the number of samples in the majority class (Kuhn et al., 2017). The two other metrics circumvent this problem, by describing two ratios that take into account how skewed a prediction is.

Metrics for Regression

To estimate how well a regression model f is able to approximate the real output y by its prediction $\hat{y}$, the *correlation coefficient (Pearson)* and the *mean absolute error (MAE)* are reported. The Pearson correlation coefficient is defined by (Clauss et al., 1974, p. 117):

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{y})^2 \sqrt{\sum_{i=1}^{n}(y_i - \bar{\hat{y}})^2}}} \tag{3.7}$$

where n is the amount of samples and $\bar{}$ denotes the mean over either the known output y or the predicted output $\hat{y}$. The MAE is defined by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.8}$$

The MAE is reported instead of the root mean squared error (RMSE) because it is easier to interpret and the error and its severance have a linear relationship in the current case. This means a prediction error of 2 is only twice as bad as a prediction error of 1 (and not worse). Therefore, the MAE is a more desirable metric for this study.

3.7.3 Data Partition – Testing Strategy

For all experiments a standard Leave-25%-speaker-out Cross Validation (CV) was applied (to end up with multiple, different subsets as proposed, for example, in Guyon et al., 2003). The four speakers recorded in each corpus location were randomly assigned to four groups, balancing the locations in the different subsets. This splitting provides the model with sufficient data ($\approx$ 480 speakers per fold) and sufficient material describing the geographic distribution of features. Furthermore, it allows the model's generalization to be tested as each speaker tested has not been encountered by the model before (Guyon et al., 2003).

Since the full set of four speakers did not exist for all locations, this resulted in slightly unbalanced sets regarding the number of speakers.

3.7.4 Evaluation of Features

As mentioned before, RFs and DTs are able to assess the quality of features, by using a measure called Variable Importance (VI). The VI is a measure that evaluates the contribution of a feature to the classification/regression problem in trees and, subsequently, in forests (Breiman, 2001). The VI is based on the impurity measure used to split nodes during the growth of the trees. An estimated VI is calculated by the resulting decrease of impurity that each variable contributes to the tree (i.e., it is based on the Gini index in classification and on the response variance in regression; Wright et al., 2015). The VI has two general problems. These are that features that have a high VI "mask" correlated features and in classification tasks the VI inherits the weakness of the Gini index, meaning that it prefers features with many outcomes over ones with few (e.g., Louppe, 2014; Nembrini et al., 2018). Despite these problems, the VI is used to evaluate the importance of features, for example, in Archer et al. (2008), Xue et al. (2006), and Liaw et al. (2002b). In line with these studies, this metric will be used in the current study to assess the quality of features and to perform a feature selection.

Two task-specific problems regarding the VI that are relevant for the current study will be discussed in the following.

First, the masking of correlated features has two drawbacks. a) it may result in a high ranking of hard-to-explain features even though easier-to-explain features would possess exactly or almost the same predictive power. This masking is a problem when it comes to interpretation. b) several highly correlated features might end up under the top-ranked features. Therefore, by no means is a minimal set of features produced. This only poses a problem for the size of the final set and not for its predictive power, as all important features that have been used during the prediction in the RF are highly ranked (even though they might be highly correlated and therefore redundant) and those features that have been discarded (because they are masked) can be left out, as they would not contribute more information than the higher ranked (correlated) features.

Second, the VI only asses the overall decrease in impurity in the tree, i.e., it is known how important a feature is to the forest, but it is unknown how it was used to split the

feature space. This missing information constitutes a problem if the features are supposed to be mapped onto a geographic space, to see if these distributions coincide with known dialectal boundaries. This drawback is ignored for experiment 1 and 2. In experiment 3, the VI is used to select the best features from a large feature set (consisting of a combination of all features of all phonemes; cf. Sec. 3.10). A DT is then trained using this subset, which allows a geographic visualization of features selected at various splits in the tree. Additionally, the VI is used to discard those features that did not contribute to the prediction model at all (cf. Sec. 3.8.6).

3.8 Experiment 1 – Binary Classification of Speakers

3.8.1 Experimental Design

Two two-class classification models are evaluated in this experiment to test the predictive power of the extracted features per phoneme. This means, that for each phoneme an individual model is trained, which tries to predict speaker origin based on a single uttered phoneme (e.g., for a phoneme with a length of 100 ms, this translates to 30 ms of information as mentioned in Sec. 3.6.2). In each direction, the corpus area is split into two halves and for each direction, an RF model is trained to distinguish speakers from the respective half. Therefore, the tasks are to assign speakers correctly to a) the North or South half of the corpus area (first model) and b) the East or West half of the corpus area (second model).

This classification will show whether the extracted features reflect the presumed regional variation contained in the speech sample and if this is sufficient to perform at least a rough division of speaker origins. If such a fairly simple two-class classification is not possible above the level of chance, the more challenging regression task is likely to fail as well. Additionally, if the two-way classification works, but a continuous estimation of speaker position does not, this could constitute a fallback, allowing at least a rough estimation of the speaker position.

The experiment in this section involves the following steps:

1. Dividing up the speakers based on their origin and assigning an appropriate label

(North/South and East/West[13]).

2. Performing a hyperparameter search on the RF for both directions separately.
3. Identifying phonemes and features that work well for each direction.
4. Detecting features that do not contribute to the prediction at all (noise features).

3.8.2 Division of Speakers

As stated above, two binary classifiers were trained. For this, the corpus was divided twice into two halves, once in the east-west and once in the north-south direction. The labels are created by splitting the corpus area at the midpoint. This point is defined by the center of gravity of all 165 recording sites' positions in the corpus (cf. horizontal and vertical line in Fig. 3.2). The midpoint is located at[14]:

- East/West (longitude): $10.41484°E$
- North/South (latitude): $50.01903°N$

The division is carried out by using a bottom-up approach independent of any top-down knowledge, this means for example, a dialectological-driven east-west and north-south separation.

North/South division: All speakers originating below or at the same altitude as the corpus midpoint were grouped in the "South" class and the remainder in the "North" class.

[13]In the remainder of this chapter, *East/West* and *North/South* are used to denote the classes of the respective region in the corpus area. In contrast to that, *east-west* and *north-south* are used to describe the respective direction (also in regression tasks).

[14]All geodetic datums in this thesis are *WGS84 geodetic datums* specified by longitude and latitude (National Imagery and Mapping Agency, 2000). "Easting" (E) describes a position in the east-west direction and is defined as a positive number that describes a position eastwards and a negative number describing a position westwards from a north-south reference line. Accordingly, it defines "Northing" (N) as a position on a north-south direction in which positive numbers describe a position northwards and a negative number a position southwards from an east-west reference line. In the following, the Greenwich meridian is used for Easting and the equator for Northing. Furthermore, if not otherwise stated, the longitude is listed before the latitude. All coordinates will be specified in decimal degrees (for more information, e.g., cf. Nilsson et al., 2004).

This division resulted in 334 speakers (52.19%) in the "South" class (159 male, 175 female) and 306 speakers (47.81%) in the "North" class (153 male, 153 female).

East/West division: Similarly, all speakers originating west of or at the same position as the corpus midpoint were grouped in the "West" class and the remainder in the "East" class. This split resulted in 340 speakers (53.125%) in the "West" class (164 male, 176 female) and 300 (46.875%) speakers in the "East" class (148 male, 152 female).

3.8.3 Results of the Random Forest Parametrizations

It has already been mentioned in Sec. 3.7 that the influence of the two main RF hyperparameters *mtry* and *ntree* on the prediction were tested. First, *mtry* was varied in three steps 27 ($\approx \sqrt{d}$), 100, and 245 ($\approx d/3$) with a fixed number of 100 trees. Second, *ntree* was varied in three steps with 100, 150, and 250 trees with a fixed $mtry = 27$. A full grid search with all combinations was not performed to save processing time.

For both hyperparameters, the resulting average accuracy over the individual phoneme models was calculated. The mean accuracies for the three different *mtry* values $\sqrt{d}$, 100, and $d/3$ (with $ntree = 100$) and the three values for *ntree* 100, 150, and 250 (with $mtry = \sqrt{d}$) can be seen in Table 3.3, where the values are presented separately for the division in each direction. It can be seen that the RFs classification results are insensitive to changes in these hyperparameters, which agrees with previous findings (e.g., Breiman, 2001; Díaz-Uriarte et al., 2006; Archer et al., 2008).

The improvement in classification accuracy was, however, statistically significant for both directions regarding higher values in *ntree* and for the North/South distinction also for *mtry* 100. Significance was estimated based on Bonferroni-corrected paired one-sided Wilcoxon-Mann-Whitney tests using the prediction accuracy of all 42 phonemes as input. For *mtry* the combinations $\sqrt{d}$ vs. 100 and 100 vs. $d/3$, and for *ntree* the combinations 100 vs. 150 and 150 vs. 250 were tested. Tests with a significance level of $d < .01$ are marked by "**".

Based on the hyperparameter tuning, the results presented in the remainder of this section describing experiment 1 are in the north-south direction for *ntree* 250 and *mtry* 100

Table 3.3: Average accuracy over all 42 phonemes in the classification task for a) *mtry* values $\sqrt{d}$, 100, and $d/3$ for 100 trained trees and b) *ntree* values 100, 150, and 250 when trained using $mtry = \sqrt{d}$. Statistically significant improvements over the next lower value is indicated by two stars ** ($p < 0.01$).

Division	**mtry**			**trees**		
	$\sqrt{d}$	100	$d/3$	100	150	250
North/South	0.6034	0.6054**	0.6053	0.6034	0.6074**	0.6112**
East/West	0.5358	0.5360	0.5365	0.5358	0.5382**	0.5400**

and in the east-west direction for *ntree* 250 and *mtry* 27 ($\sqrt{d}$).

3.8.4 Classification Results – North/South

Table 3.4: Classification accuracy, precision, recall, and NIR of the five top-ranking phonemes for North/South classification; ordered by accuracy and rounded to four decimals for both classification tasks.

Phoneme	**Accuracy**	**Precision**	**Recall**	**NIR**
/z/	0.7037	0.6923	0.6466	0.5377
/øː/	0.6898	0.6696	0.5958	0.5557
/ʏ/	0.6612	0.6384	0.3391	0.6028
/i/	0.6474	0.6499	0.7043	0.5223
/aʊ/	0.6455	0.6471	0.6867	0.5153

Table 3.4 shows the results for the five best phonemes, ranked by their accuracy. It is worth mentioning that all phonemes (including the phonemes not shown in Table 3.4) predict the correct geographic class above the level of chance. The worst accuracy of 0.5689 is achieved with the phoneme /o/.

Table 3.5 lists the best features for the phonemes /z/ and /øː/. It can be seen that,

Table 3.5: The top ten features for the best-performing phonemes /z/ and /øː/, ranked by VI. In the case that a feature is a vector, its index is given in parentheses, starting at 0. As a reminder, a list of acronyms can be found at the beginning of this thesis.

/z/	/øː/
VU	MFCC (7)
VC (0)	MFCC (8)
AS (13)	AS (13)
AS (14)	STS (61)
AS (2)	MFCC (5)
ZCR	MFCC (3)
SE	AS Rfilt (10)
MCR	AS (10)
MFCC (8)	AS Rfilt (9)Δ
AS (16)	AS (10) Δ

apart from features describing the periodicity of a phoneme, like voicing, MCR, ZCR, and SE, MFCC and AS features are the most prominent features.

Once again, it is worth noting: the voicing probability that can be linked to a devoicing of /z/ in southern varieties does not necessarily mean that the voicing probability extracted from the speech signal only captures strongly voiced speech segments. Due to the large amount of data available, as stated before, a manual auditory validation of the assumed connection between extracted acoustic features and signal was not performed.

Interestingly, only two short-time functionals are listed for /øː/: the Δ of AS coefficient (10) and the Δ of RASTA-filtered AS coefficient (9). This trend is continued within the top 50 features, where for /z/ only ten (20%) and for /øː/ only 13 features (26%) are Δ or ΔΔ features. When taking all phonemes into account, only around $\approx 33\%$ of the top 50 features of all phonemes are Δ and ΔΔ features. This might be due to three reasons: a) these features do not model regional variation well, b) they are masked by other features that are ranked higher, or c) the Δ and ΔΔ features are more influenced by context due

to coarticulation, which does not generalize well.

The features VU and VC (0) of phoneme /z/ are a good example of the masking effect of the VI mentioned earlier. According to the VI, they are the best and second best feature for prediction. However, they are almost perfectly correlated with $R = 0.9999997$ and both are identical except for 17 realized /z/ (17 differences for 46566 realized /z/; the absolute differences in those 17 cases cumulate to $6.85 \cdot 10^{-5}$).

Phoneme /z/: The best features reported are VC, VU, measures of the periodicity of the signal (like ZCR and MCR), and a measure of the periodicity of the spectrum in case of SE (cf. Table 3.5, left column). These can be linked to the often reported devoicing of German standard non-final /z/ in southern varieties of German (e.g., König, 1989, p. 93; Barbour et al., 1990, p. 156; Wängler, 1967, p. 143) in many positions[15]. [z̥] substitutes an /s/-like sound that is pronounced more weakly, instead of the stronger pronunciation of the actual /s/; the aforementioned literature agrees that the lenis character of /z/ is still present in the pronunciation of the devoiced form in the southern varieties.

A good separation can be achieved between northern and southern speakers of the corpus area using the feature VU of phoneme /z/ alone. The distribution over the corpus area of this feature is shown in Fig. 3.4, in which a clear separation can be seen between the North and the South. In the plot, each value is the average over multiple realizations of /z/ for a speaker and has then been normalized to a range between 0 and 1 using the 5% and 95% quantiles (so as to make them more robust against outliers). The colors are taken from a perceptually balanced color-scale (Moreland, 2009).

The AS features describe the frequency bands 139.62 Hz−312.76 Hz (AS (2)), 1644.00 Hz−2127.37 Hz (AS (13)), 1873.31 Hz−2403.96 Hz (AS (14)), 2401.42 Hz−3040.97 Hz (AS (16)), and 2704.83 Hz−3406.94 Hz (AS (17)). The lower band of AS (2) especially, is a location in the spectrum, where a large difference between [z] and [z̥] is to be expected. In this band, the voice bar of voiced /z/ should lead to higher energies in northern speakers, which is

[15]Phoneme /z/ is voiced in standard German word-initial in front of a vowel, intervocalic, between /m, n, ŋ, l, r/ and vowel, and after /p, b, d, t, g, k/ in epentheses and in suffixes *-sam*, *-sal*, and *-sel* (König, 1989, p. 93)

Figure 3.4: A map showing the distribution of feature VU for the phoneme /z/. The values are averaged over all realizations of a speaker and then normalized to the range between 0 and 1 using the 5% and 95% quantiles so as to be more robust against outliers. Blue colored circles indicate low values for the voicing probability (close to 0), red colored circles indicate high values for voicing probability (close to 1), and gray colored circles indicate values in the middle of the scale (around 0.5).

indeed the case (cf. Fig. A.1). As expected, in the other three bands, more energy can be found in the southern speakers due to more frication in [z̥] (cf. Fig. A.1).

MFCC (8) translates to a cross-correlation of a cosine with $3\frac{1}{2}$ cycles over the spectrum. Having this many peaks and troughs, this feature is a bit more difficult to explain. Especially because the spectral envelop looks similar for both northern and southern speakers (cf. Fig. A.2). However, the majority of the northern speakers start the first cycle below 0, the majority of the southern speakers above 0. As the feature values are closer to 0 for the southern speakers, the spectrum is flatter for them. The peak of the first cycle is at around 250 Hz for the northern speakers. This correlation could be once again interpreted as an indicator of the voice bar. The rest of the peaks over the spectrum cannot be fully explained (cf. Fig. A.2).

Phoneme /øː/: The second best phoneme for distinguishing the North from the South is /øː/ (cf. Table 3.5, right column). One reason for this could be the realization of /øː/ as an [øː]-like sound in the North of the corpus area, although the realization is more [eː]-like in Bavarian and Swiss varieties (Landesbibliothek, 2013; Kleber, personal communication, 2018). This assumption is supported by the fact that many features occupy spectral bands that are normally occupied by the second formant of vowels (cf. Fig. 3.5b).

The second formant (F2) describes the change of the horizontal tongue position and lip rounding (vocal tract shape changes are not independent, cf. *Formantverschieber*[16] in Tillmann et al., 1980, p. 262). Assuming the formant values describe a difference between [øː]-like and [eː]-like realizations, the decrease in F2 mostly describes the presence of more lip rounding in the vowel /øː/. It could also describe a difference in the retraction of the tongue. Features that characterize the energy present in those bands appear in different forms, with or without Rfilt.

The top features describe the following frequency bands: 913.70 Hz − 1246.46 Hz (AS Rfilt (9)), 1071.56 Hz − 1436.88 Hz (AS (10), AS Rfilt (10), and AS Rfilt (10) Δ), and 1644.00 Hz − 2127.37 Hz (AS (13)). This is supported by the fact that the northern speakers have more energy in the lower bands AS (9) and AS (10), and the southern speakers more

[16] *Formantverschieber* roughly translates to "formant shifter".

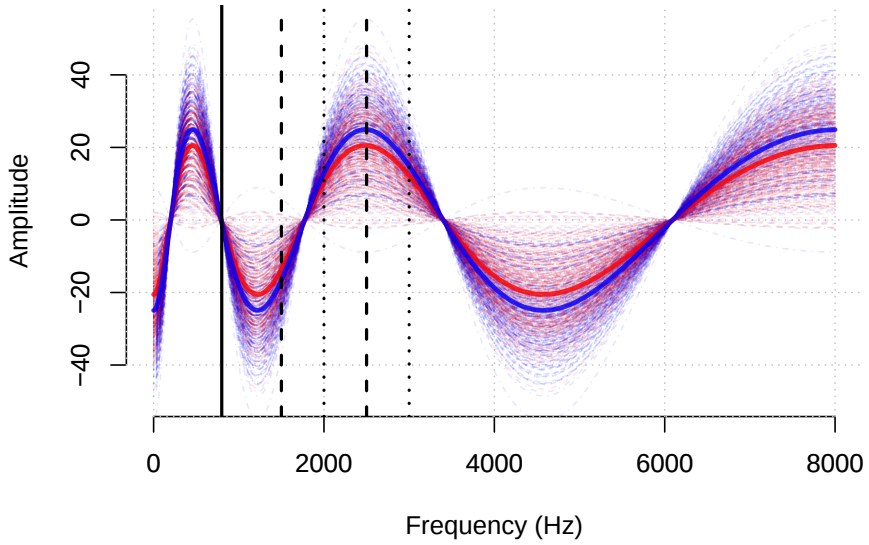
in the upper band AS (13). This could mean that southern speakers move the F2 closer to an /eː/-like target (cf. Fig. 3.5b; cf. boxplots of feature values in Fig. A.3).

The feature STS (61) has a similar frequency range to AS (13). It describes the spectrum between 1812.33 Hz and 1920.095 Hz, which is the band around the center frequency of AS (13) at 1885.69 Hz. Therefore, it could describe the energy in the second formant of [eː], where one would expect to be more energy present for the southern speakers, which is the case (cf. boxplots of feature values in Fig. A.3).

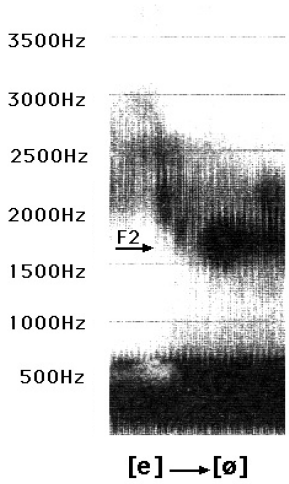
A resynthesis of the MFCC (5) can be seen in Fig. 3.5a. Vertical lines are plotted at the points distinguishing /øː/ from /eː/ at 800 Hz (solid line), 1500 Hz (left dashed line), 2500 Hz (right dashed line), 2000 Hz (left dotted line), and 3000 Hz (right dotted line). Between the solid and the left dashed line, less energy for the southern speakers would be expected due to the [eː]-like pronunciation. Between the two dashed lines more energy for the northern speakers ([øː]), and between the two dotted lines more energy for the southern speakers ([eː]) is to be expected.

The first and third assumption hold true for MFCC (5). The second assumption lies close to the turning point of the cosine wave, where it has only little influence on the amplitude of the cosine wave. When looking at MFCC (2) – not in the top ten list – describing the curvature of the spectrum, it can be seen that the spectral envelope of the northern speakers is flatter across the spectrum than for southern speakers. This would agree with an /eː/-like pronunciation of the southern speakers, since in these sounds the change in energy across the spectrum is supposed to be greater than for [øː]-like pronunciations (cf. 3.5b)

Another problem that arises when interpreting MFCCs is that they do not only describe frequencies of interest in the current phoneme-class (in the current case vowels for which the range between $0\,\text{Hz} - 4000\,\text{Hz}$ is of special interest) but the whole spectrum. Nevertheless, MFCCs have often been shown to model all kinds of speaker variability and para-linguistic traits, such as emotion recognition (e.g., Sato et al., 2007; Schuller et al., 2009), speaker recognition (e.g., Murty et al., 2006; Tiwari, 2010), and dialect classification (cf. Sec. 3.3).



(a) Resynthesis of the coefficient MFCC (5) for phoneme /øː/ with lines at 800 Hz (solid line), 1500 Hz (left dashed line), 2500 Hz (right dashed line), 2000 Hz (left dotted line), and 3000 Hz (right dotted line). The lines indicate resynthesized spectral envelopes of speakers from the North (plotted in red) and the South (plotted in blue). The two thick lines are the resynthesis of the averaged MFCC over all speakers for the respective group.



(b) Example spectrogram of [e] vs. [ø] (Machelett, 1996).

Figure 3.5: Resynthesis of the MFCC coefficient five (left) for North/South classification and the spectrogram of an example of the two phonemes it might distinguish (right).

3.8.5 Classification Results – East-West

The results of classification accuracy, precision, recall, and NIR are summarized in Table 3.6 for the top five phonemes in the east-west direction. It can be seen that even the five best phonemes do not perform well above the level of chance. Overall, the classification accuracies of all shown phonemes lie close together and no phoneme reaches an accuracy above 58%.

Table 3.6: Classification accuracy, precision, recall, and NIR of the five top-ranking phonemes for East/West classification; ranked by accuracy and rounded to four decimal places for both classification tasks.

Phoneme	Accuracy	Precision	Recall	NIR
/øː/	0.5791	0.5891	0.7498	0.5443
/z/	0.5754	0.5829	0.6929	0.5287
/ɛː/	0.5718	0.5768	0.7007	0.5264
/u/	0.5667	0.5675	0.5700	0.5013
/ç/	0.5593	0.5663	0.7228	0.5306

The best features for the once again best phonemes /øː/ and /z/ are shown in Table 3.7. Not even one short-time functional is present under the top ten features for both phonemes. This is a trend that continues, for phoneme /øː/ only eleven Δ and $\Delta\Delta$ features are under the top 50 features and for phoneme /z/ there are only four $\Delta\Delta$ features present.

Phoneme /øː/: The features of /øː/ are partially identical to the North/South distinction case (cf. 3.7, left column). The fact that the used features are mostly the same might stem from the fact that the straight vertical line used for separating East and West does not reflect the dialectological reality and no better distinction between “East” and “West” can be made using other features from different phonemes.

The reappearing features are MFCC (3), MFCC (7), MFCC (8), AS (10) (1071.56 Hz – 1436.88 Hz), AS (13) (1644.00 Hz – 2127.37 Hz), and STS (61) (1812.33 Hz – 1920.095 Hz).

Table 3.7: The top ten features for the best-performing phonemes /øː/ and /z/, ranked by VI for the East/West classification. If a feature is a vector its index is given in parentheses starting at 0.

/øː/	/z/
MFCC (8)	MFCC (8)
MFCC (3)	AS (20)
MFCC (10)	LSP (0)
AS (13)	AS (17)
STS (88)	AS (18)
STS (87)	STS (76)
AS (10)	MFCC (3)
MFCC (7)	AS (19)
SV	MFCC (1)
STS (61)	STS (75)

That the separation is partially an artifact of the crude East/West separation is supported by the actual feature values. In them, the East appears to behave similarly to the South and the North similar to the West (cf. boxplots of feature values in Fig. A.5).

The other features do not show a clear east-west distinction when their averages are plotted over a German map. STS (87) (8137.08 Hz − 8620.93 Hz), STS (88) (8620.93 Hz − 9133.56 Hz), and SV only distinguish particular parts on the map. This suggests that the values overlap to a great extent, which is the case (cf. feature values in Fig. A.6). The fact that no clear East/West separation can be observed in the plotted feature values is taken as evidence that distinguishing East from West is not as trivial as North from South when using only a single phoneme.

Phoneme /z/: That the phoneme /z/ is known to be devoiced in the southern German varieties was already mentioned in Sec. 3.8.4. Based on the data-driven axis-parallel division of the corpus area the method uses a straight line for North/South separation. This

separation is unlikely to reflect the dialectological reality. A more realistic separating line would be a diagonal, curved one. This mismatch could be the reason why Eastern speakers behave like speakers from the South and Western speakers like ones from the North. Therefore, similar features based on the same variation are used to distinguish the (data-driven) East and West halves of Germany. One feature that supports this hypothesis is MFCC (8) (cf. 3.7, right column) as it appears under the top ten features in the North/South distinction as well. Even though this feature is hard to explain, the existence in both top ten feature lists means that it can be assumed that the same information is used.

The feature values for AS (17) (2704.83 Hz – 3406.94 Hz), AS (18) (3037.91 Hz – 3808.71 Hz), AS (19) (3403.58 Hz – 4249.79 Hz), and AS (20) (3805.03 Hz – 4734.02 Hz) further support this claim. Like for the southern speakers in the North/South distinction, the feature values are higher in this band for speakers originating from the East. This behavior is expected for a devoiced /z/.

Additional evidence is given by the resynthesis of the feature MFCC (1) as speakers from the West of the corpus area tend to devoice more. In the resynthesis, it can be seen that the slope for the speakers from the East rises (more energy in higher bands), whereas the speakers from the West of the corpus area have a flatter envelope (cf. Fig. A.8).

The LSP (0) is an interesting feature as it did not appear in the North/South distinction (cf. Secs. 3.6.3 and 3.6.3). LSP (0) has lower values for North and West, which might be an indicator of the existence of a voice bar since the energy in the envelope would need to be reconstructed there. The fact that this difference is due to the voice bar being present or not, is supported by higher values for the East German speakers, just like the higher ones for the southern speakers. (cf. Fig. A.9).

3.8.6 Noise Features

During the experiment, 81 features showed a VI of 0.0 in all phonemes. Those features are (vector index in parentheses starting at 0):

- Linear Predictive Coding (LPC) (3,4,5,6,7) (*five features*)
- Semi-Tone Spectrum (STS) (0-7,9-11,13,14,16,18,19,21,23,26,30,94,95) (*22 features*)

including the respective Δ and ΔΔ features (*54 features*). Due to the fact that they do not contribute to the prediction, these features will be excluded from the feature sets of the subsequent experiments to save processing time.

3.8.7 Discussion

It could be shown that many phonemes can be used to distinguish speakers from different parts of Germany. All phoneme types available in the corpus can be used to classify North/South above the level of chance, and 31 (of 42) phonemes can be used to classify East/West above the level of chance.

The best performance achieved in any direction was 70.37% for the phoneme /z/. The achieved accuracy is around 11% worse than in the previous experiment reported in Kisler et al. (2018b). This is because for all speakers the majority vote of the predictions for all of his/her realized phonemes was used to decide on the final prediction in Kisler et al. (2018b), whereas in the current study only a single uttered phoneme was used. This changes the results for two reasons. First, not every single uttered phoneme /z/ is devoiced, not even in the southern varieties. This is because, depending on the context, not every /z/ is supposed to be devoiced (even though those cases are in the absolute minority). Second, due to intra-speaker variability in spontaneous speech, not all speakers always fully devoice all /z/.

Another reason affecting the performance negatively is the division of the respective halves using a straight line, something that also negatively influenced the results in Kisler et al. (2018b). This shape does not reflect dialectological reality, where a diagonal, curved line would more accurately portray reality. Using a dialectological motivated line to separate the two halves would improve the results, though would contradict the pursued bottom-up approach.

The results from the current study suggest that even a small subset of phoneme types could be sufficient for a regression analysis. Only relying on a few phonemes might enable applications, for example, in an ASR system to be able to estimate a speaker's origin from his or her first few spoken words. Furthermore, good estimation performance based

on a small number of phonemes seems to be more likely for North/South than for East-/West. This is because the accuracy for the East/West distinction was worse than for the North/South distinction. One would expect classification schemes based on linguistic features such as dialectal word forms or phonemic n-grams (Stadtschnitzer et al., 2014) are expected to require much more input data from the target speaker then used in the current case.

Next to only using a few phonemes, it seems that it might be possible to base the localization on only a few features for the North/South classification. Using only a few different features seems questionable for the east-west direction as a) the results are much worse and b) the VI is more equally spread over the features (cf. Fig. A.10[17]). The low VI might indicate that more features need to be combined to achieve a result that still does not reach the performance of the North/South distinction.

Fig. 3.6 shows the relationship between the accuracy achieved in the North/South and the East/West classification task for each phoneme. It is interesting to see a strong correlation between the accuracy in both directions ($R = 0.6112$). This correlation could be partially due to the above mentioned artifact caused by the crude separation of the respective halves with two straight, axis parallel lines. However, different features are used for the two different directions, which might be an indicator that the entire performance is not a result of only this artifact.

When regarding the VI of features, it is important to keep in mind that features are always combined with other features and, by definition of the RF, features are combined in multiple different ways due to random selection at each split. Nevertheless, the steep decrease in importance in the north-south dimension is evidence for features that much better model variation than others (cf. Fig. A.10).

A total of 81 ($\approx 11\%$ of the set of 737) features in the openSMILE feature set were found unsuitable for the classification task. In further tests these features can be omitted to save processing time.

[17] For comparability of the VI values they are compared to forests grown with the same *mtry* value ($mtry = 100$) as this parameter controls the models' complexity; in all other results for the east-west dimension *mtry* was set to $\sqrt{d}$.

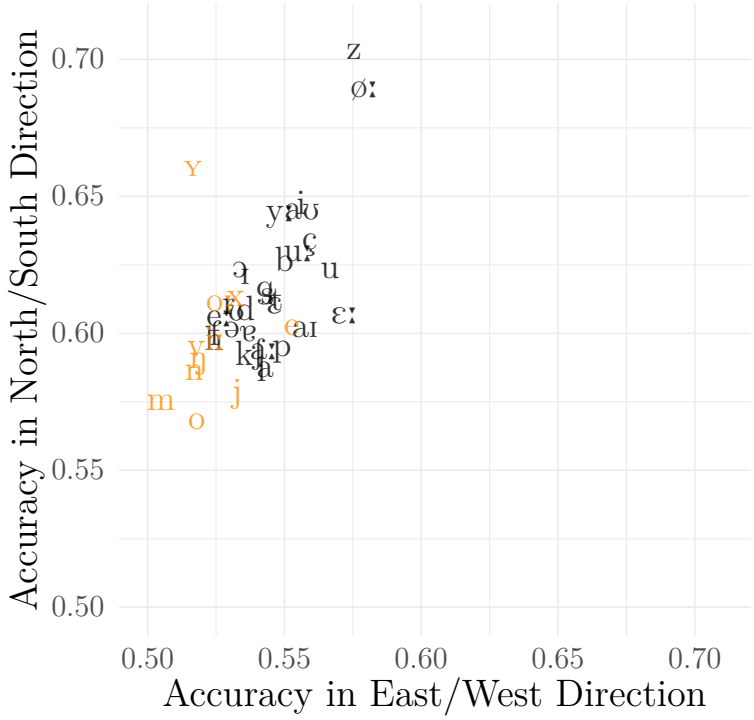


Figure 3.6: A plot of phoneme-wise accuracy for both directions (east-west on the x-axis, north-south on the y-axis). For the North/South classification all phonemes were above NIR. These phonemes that were above NIR in the East/West classification are drawn in black, those below NIR in orange.

3.9 Experiment 2 – Regression of Speaker Location

3.9.1 Experimental Design

The previous experiment showed that extracted features contain information that relates to geographic variation and therefore enables the prediction of a speakers' origin above the level of chance. The experiment in the following section builds on the results of the intermediate step that was experiment 1 and answers the questions: *Is it possible to continuously estimate speaker origin in a geographic space based on a single phoneme extracted from a speech sample?* And if so: *How exact is this localization?*.

The experiment will be performed in an analogous fashion to experiment 1, the difference being that the RF predicts continuous positions instead of class membership. For the current experiment, the original dataset introduced in Sec. 3.6 will be used, minus the 81 noise features listed in Sec. 3.8.6. Removing the noise features leads to a feature vector of dimensionality $d = 656$.

Conducting the experiment involves the following steps:

1. Select a baseline for comparison.
2. Performing a hyperparameter search on the RF for both directions separately.
3. Identifying phonemes and features that work well for each direction.

3.9.2 Selection of a Baseline

As mentioned before, to my knowledge regression analysis of speaker positions has not been previously examined. Missing previous research, in turn, means that no baseline exists to which a model can be compared. To circumvent this problem in the following experiments, a *null model* is used instead. This model has no predictors and only returns the mean of the respective variable (James et al., 2014, p. 205).

Two possible null models are described in the following, one of which is a geometric and one a data-driven null model. The geometric null model uses the center of gravity of the desired speech area and returns this as a prediction. The data-driven null model, on the other hand, returns the center of gravity of the recording sites. Both of these models

are conservative and, intuitively, will not result in good predictions of speaker origins.

The difference in the resulting baseline error between the two models is small (east-west: 5.03 km; north-south: 1.00 km). Furthermore, the basic assumption is (and has to be for any corpus analysis) that the corpus is in fact a good representation of the real world regarding the phenomena of interest. This means in the current case that a) the material is sufficiently spread over the German-speaking area and that b) it captures relevant regional variation necessary for a speaker localization. Therefore, it seems more consistent to use the data-driven null model in the current case.

The corpus midpoint was already used in the classification experiments so as to be able to group speakers into the four classes "North", "South", "East", and "West" (east-west: $50.01903°E$; north-south: $10.41484°N$; cf. Sec. 3.8.2), and is shown as a black cross in Fig. 3.7. The null model results in an error of 151.44 km (2.1191°) for longitude and 210.89 km (1.8960°) for latitude. To visualize this error in the German-speaking geographic space, the error is shown as an ellipse in Fig. 3.7. This ellipse describes the average uncertainty that has to be taken into account when any point is predicted. It is worth noting that this is a hypothetical error, as the null model only returns a single point.

Figure 3.7: The midpoint of the GT corpus plotted on a German map (black cross) together with the baseline error for the null model (black dashed ellipse).

3.9.3 Results RF Parametrization

As in the previous experiments, different RF parametrizations were tested. The results can be seen in Table 3.8. For *mtry* the values $27 \approx \sqrt{d}$, 100, and $218 \approx d/3$ were tested for a fixed number of 100 trees and for *ntree* the values 100, 150, and 250 were, once again tested, for a fixed *mtry* of $d/3$. Statistical significance was estimated using Bonferroni-corrected paired one-sided Wilcoxon-Mann-Whitney tests that used the prediction accuracy of all 42 phonemes as input. For *mtry* the combinations $\sqrt{d}$ vs. 100 and 100 vs. $d/3$ were tested. For *ntree* 100 vs. 150 and 150 vs. 250 were tested. The significant tests are marked by "**", where the significance level in those cases was $d < 0.01$. A full grid search with all possible combinations was again not performed to save processing time.

Similar to the results in the classification experiment, certain combinations of values had a significantly better result, even though the differences were marginal. Statistically significant results were achieved for higher settings of *mtry* in the north-south direction, where the values 100 and $d/3$ were significantly better than the two other values. Similarly, in both dimensions more trees resulted in significantly better results as well. Therefore, the hyperparameters of the best result will be used in the remainder of this section. This means for longitude the hyperparameters $mtry = \sqrt{d}$ and $ntree = 250$ and for latitude $mtry = d/3$ and $ntree = 250$.

Table 3.8: MAE for different parametrizations of the RF. The *mtry* values tested were $\sqrt{d}$, 100, and $d/3$ (for a fixed number of 100 trees) and the *ntree* values 100, 150, and 250 (for a fixed number of *mtry* $\sqrt{d}$). The results that are significantly better than the neighboring lower value are marked by two stars ** ($p < 0.01$). The unit for all values is kilometers (km).

Direction	**mtry**			**trees**		
	$\sqrt{d}$	100	$d/3$	100	150	250
North-south (Latitude)	206.53	205.86**	205.61**	205.61	205.36**	205.17**
East-west (Longitude)	147.85	147.98	148.05	148.05	147.82**	147.64**

3.9.4 Regression Results – North-South Direction

The five best-performing phonemes are listed in Table 3.9, ranked according to their MAE. As in the classification experiment, the best prediction was possible for the phonemes /z/ and /øː/. Altogether, for 34 phonemes a prediction was possible that is better than the baseline defined in Sec. 3.9.2. The correlation turns out to be weak to moderate. Together with a small improvement over the baseline, this means the prediction generally points in the right direction, but the model does only predict positions that are close to the midpoint.

Table 3.9: MAE and Correlation (Cor) of prediction and real values for the five best-performing phonemes' in the north-south direction.

Phoneme	**MAE**	**Cor**
/z/	183.70 km	0.4593
/øː/	189.22 km	0.4261
/ʏ/	190.46 km	0.3054
/ç/	197.52 km	0.3134
/x/	198.80 km	0.3012

An example of this is the moderate correlation of $R = 0.4535$ for the prediction of the best phoneme /z/. Despite this correlation, the improvement of the prediction over the null model is only 26.69 km (0.2399°). This translates to a relative improvement over the baseline of $\approx 12.65\%$.

For eight phonemes the predicted values show a worse MAE than the baseline. This is somewhat surprising as in experiment 1 all phonemes could be used to estimate the North/South half above the level of chance. The phonemes that resulted in an MAE worse than the baseline are, in ascending order of their MAE, /aɪ, f, h, j, m, o, p, v/. The worst MAE was achieved by phoneme /v/ with 213.94 km (1.9235°) and the worst correlation was achieved by phoneme /m/ with $R = 0.1592$.

The ten best features are shown in Table 3.10 for the two best-performing phonemes

/z/ and /øː/. It is interesting to note that nine of the ten features are already present in the top ten features for North/South classification (cf. 3.8.4). Furthermore, the features for both phonemes appear in almost the same order as in the classification task, when ranked according to the achieved VI.

Table 3.10: The top ten features for the best-performing phonemes /z/ and /øː/, ranked by VI for the north-south direction. If the feature is a vector, the index is given in parentheses starting at 0.

/z/	/øː/
VU	MFCC (7)
VC (0)	MFCC (8)
AS (14)	AS (13)
AS (13)	STS (61)
SE	MFCC (5)
AS (16)	AS (10)
AS (2)	AS Rfilt (10)
MCR	MFCC (3)
AS (15)	AS (10) Δ
ZCR	AS Rfilt (11)

Phoneme /z/: AS (15) is the only feature that appears in the top ten in the regression task exclusively, i.e., was not present in the classification task. This feature describes the energy in the band between 2125.05 Hz and 2707.61 Hz. In this region more spectral energy would be expected in [z̪]-like than in [z]-like realizations, due to more frication. Therefore, since there is more energy in realizations of southern speakers, this might mean that it once again describes devoicing (cf. Fig. A.11). It replaces the nearly uncorrelated ($R = 0.02663$) MFCC (8) in the top ten.

Phoneme /øː/: AS Rfilt(11) is the only feature for phoneme /øː/ that appears under the top ten features of the regression task that had not been present in the classification task. It describes the energy in the frequency band between 1244.87 Hz and 1645.92 Hz, where, according to Fig. 3.5b, less energy is expected for [eː]-like realizations. This holds true for the southern speakers (cf. Fig. A.12). It replaces the strongly correlated ($R = 0.6121$) AS Rfilt (9).

3.9.5 Regression Results – East-West

The five best phonemes in the east-west direction are listed in Table 3.11, sorted according to their MAE. For 41 phonemes a prediction was possible that surpassed the baseline defined in Sec. 3.9.2. Only the model built for phoneme /ʃ/ outputs a prediction worse than the baseline, with an MAE of 154.65 km (2.1360°).

Table 3.11: MAE and Correlation (Cor) of prediction and real values for the five best-performing phonemes in the east-west direction.

Phoneme	MAE	Cor
/z/	141.99 km	0.1997
/ɛː/	143.11 km	0.2092
/u/	143.22 km	0.1187
/øː/	143.28 km	0.1849
/x/	144.20 km	0.0863

That so many phonemes can be used to predict the east-west position of the speaker origin above the baseline is surprising. Especially, regarding the weak correlation achieved by the top five phonemes and the fact that fewer phonemes were suitable to predict the East/West classes correctly above the level of chance in experiment 1. For example, the phoneme /x/ only shows a weak correlation of $R = 0.0863$. And even the best phoneme under the top five only achieved a correlation of $R = 0.2092$. The worst correlation

$R = 0.0002$ was achieved by the phoneme /v/.

The achieved improvement of the MAE over the baseline with the best-performing phoneme /z/ is only 9.45 km (0.132284°). This is a relative improvement of ≈ 6.24%, which is around half of the improvement of the regression in a north-south direction.

The top features according to their VI can be seen in Table 3.12, where once again no Δ or ΔΔ features are present.

Table 3.12: The top ten features for the best-performing phonemes /z/ and /ɛː/, ranked by VI. If a feature is a vector, its index is given in parentheses starting at 0.

/z/	/ɛː/
MFCC (1)	MFCC (3)
LSP (0)	AS (9)
LSP (4)	MFCC (6)
MFCC (8)	AS (10)
AS Rfilt (25)	LSP (1)
AS Rfilt (22)	AS (8)
AS (18)	MFCC (4)
AS (19)	MFCC (7)
AS Rfilt (23)	Duration
AS Rfilt (20)	MFCC (5)

Phoneme /z/: Here six of the ten top features of to the East/West classification in experiment 1 reappear: MFCC (1), MFCC (8), LSP (0), AS (18), AS (19) and AS (20), where the last feature did appear in its RASTA-filtered variant in the current regression experiment.

Two features that do not reappear are STS (75) (4068.54 Hz − 4310.47 Hz) and STS (76) (4310.47 Hz − 4566.78 Hz). However, they have a similar frequency range as feature AS Rfilt (20) (3805.03 Hz − 4734.02 Hz), which was already present in the non-RASTA-filtered version before. It is possible that they have been dropped as they cannot provide more

information in the regression task as they describe the same change in frequency in the east-west direction. When looking at the feature values of both STS features, it is apparent that they show a similar distribution with regards to their change from East to West as AS Rfilt (20) (cf. Fig. A.13 and cf. Fig. A.14).

The fact that similar features are chosen when compared to the East/West classification is taken as evidence that a similar variation is modeled. In the East/West classification, it was assumed that the straight-line separating the two classes was the reason why the phoneme /z/ was chosen (as regional variation does not occur axis-parallel). Therefore, a change happening in the north-south direction might also be able to predict speaker origins in the east-west direction.

The energy band of AS Rfilt (20) would also be influenced by devoicing due to more frication (expected to be higher for devoiced phones in that band). When plotting these features (averaged over multiple realizations for one speaker) no clear east-west distinction can be seen (cf. Fig. 3.8). The change seems to happen mostly in the South and South/East area and is generally not as widespread as, for example, shown in Fig. 3.4. Furthermore, it can be seen that many single speakers behave differently compared to speakers from the same region and recording site.

The newly added features AS Rfilt (22) (4729.57 Hz – 5849.21 Hz), AS Rfilt (23) (5260.73 Hz – 6489.91 Hz), and AS Rfilt (25) (6484.03 Hz – 7965.46 Hz) describe the upper end of the frequency spectrum. In these bands, more energy would be expected in speakers who realize /z/ as [z̥]. When looking at the actual feature values, it can be seen that the further East, the more energy is present in those bands (cf. Fig. A.14). This is taken as evidence that these features capture the devoicing of /z/ in southern speakers.

Features that only show a local regional variation (e.g., only in the south-east area in Austria) might be the reason why more features need to be combined in the east-west direction to make a prediction. This is supported by the lower and more equally spread VI in the top features (behaving similarly to the VI in the classification task, cf. Sec. 3.8.7).

Phoneme /ɛː/: It is believed that northern speakers produce the /ɛː/ like [eː] (Wängler, 1967, p. 100), while southern speakers produce it as a more open vowel (Wängler, 1967, p.

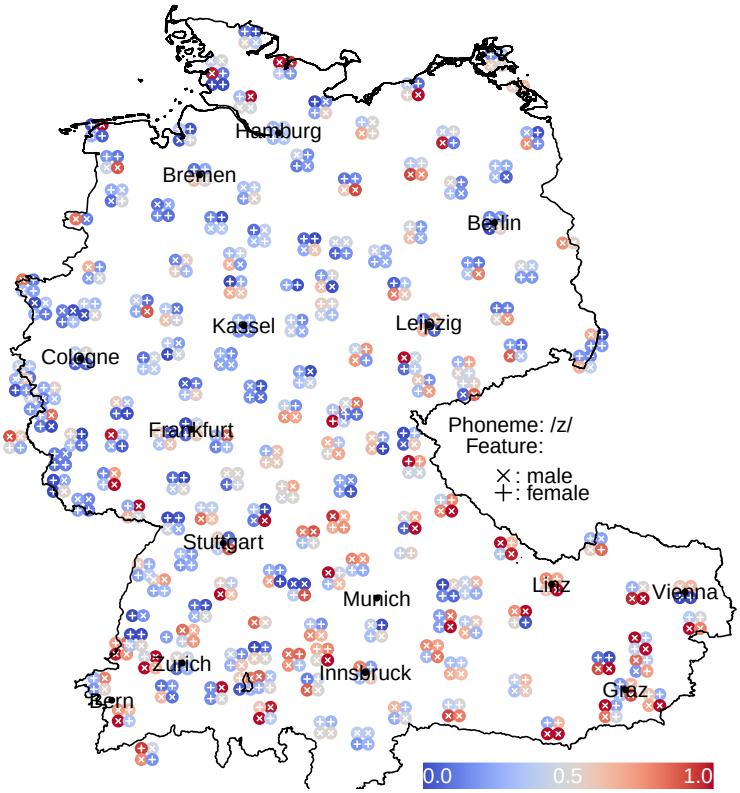


Figure 3.8: A map showing the distribution of feature AS Rfilt (20) (3805.03 Hz – 4734.02 Hz) for the phoneme /z/. The values are averaged over all realizations of a speaker and then normalized between 0 and 1 using the 5% and 95% quantiles to be more robust against outliers. Blue colored circles indicate low values for the energy band, red colored circles indicate high values for the band, and gray colored circles indicate values in the middle of the scale.

99), leading to [æː]-like and [aː]-like realizations. Once again, assuming that the variation carries over to the east-west direction, due to the non-axis parallel nature of variation, speakers from the East are supposed to behave like speakers from the South and speakers from the West, like speakers from the North.

In the energy bands AS (8) (769.90 Hz − 1073.01 Hz), AS (9) (913.70 Hz − 1246.46 Hz) and AS (10) (1071.56 Hz − 1436.88 Hz) more energy would be expected in those bands in [ɛː] compared to [eː] (rising F1 and sinking F2 for /ɛː/); and even more energy would be expected in those bands in realizations closer to [aː] when compared to [ɛː] (rising F1 and sinking F2 for /aː/). All three features describe the area where these front vowels have differences in F1 and F2 (Machelett, 1996). Based on this tripartite variation, the further East a speaker originates from, the more energy would be expected in those bands. When grouping speakers into five bins along the longitude axis, this general trend can be observed in the data. However, the westmost group shows values that are even higher than those in the eastmost group, which would suggest a /aː/-like realization (cf. Fig. A.15).

A feature that does not appear in other contexts is duration. Here the duration of the phoneme /ɛː/ is generally longer in the West than in the East. However, in the current case the speakers originating in the second westmost group, exhibit a longer duration than the speakers from the westmost group (cf. Fig. A.16).

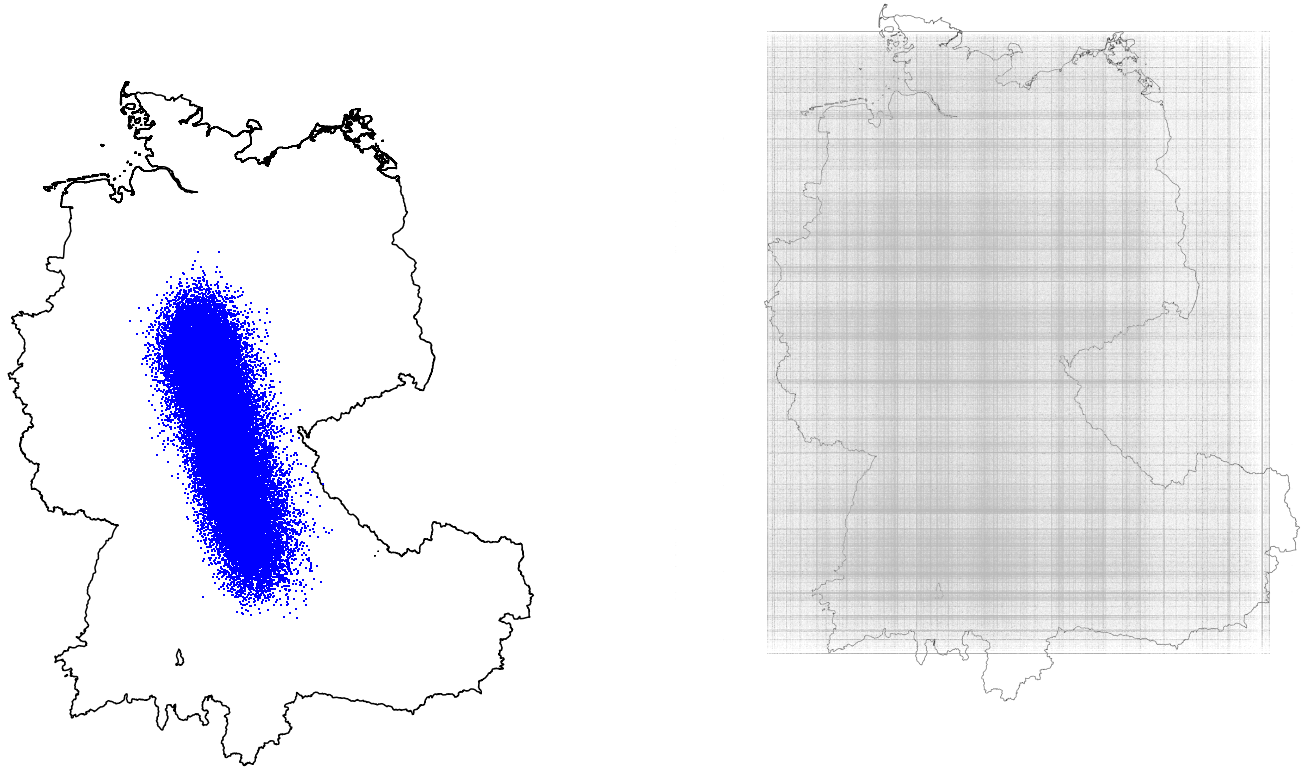
Unfortunately, /ɛː/ is also part of the canonic form of the only used hesitation marker <*äh*> (for the transcription conventions, cf. *Projekt Deutsch heute Orthografische Verschriftlichung gesprochener Sprache (Interviews und Map Tasks) KONVENTIONEN* 2015). That is, even if a hesitation is pronounced differently to its canonic form [ɛːh], it is transcribed as <äh>. Ergo, many examples can be found in which it marks parts of the speech signal pronounced like [aːm, [hm], [ɛː], etc. Hence, this phoneme contains a lot of noise with regards to its realizations. Despite this problem, based on the Leave-25%-Speaker-out CV testing strategy, each model is assessed using unseen speakers (mutually exclusive training and test set). Therefore, a model can only perform well if it indeed models regional variation. It would, however, be possible for hesitation markers to be used systematically differently across the corpus area.

3.9.6 Discussion of Regression Results

The phoneme /z/ has the smallest MAE in both directions. Even though different features were used to calculate the predictions for the two different directions (east-west and north-south), this is somewhat surprising, as a clear trend in differing feature values can only be seen for the north-south direction, and the trend is far less visible for the east-west direction. To visualize how well a speaker's position is predicted, all predictions are plotted on a map of Germany in Fig. 3.9a. The RF predictions are not spread across the map, but are instead located in the middle (blue points). This is especially true for the spread of predictions in the east-west direction. These distributions raise the question: do the trained RFs predict values close to the maximum/minimum values, i.e., close to the border of the German-speaking area?

To answer this question, Fig. 3.9b shows the predictions of the individual trees as gray dots. In order to plot the predictions in the x- and y-dimensions simultaneously, the predictions of the trees with the same number in the east-west (x) and the north-south direction (y) are combined (the first tree for longitude is combined with the first tree for latitude, the second tree for longitude is combined with the second tree for latitude, etc.). This arbitrary combination of predictions is calculated as varying the dots along both axes at the same time improves visualization. Otherwise, all values would have to be plotted on a line. As the predictions are spread over the complete corpus area, this arbitrary combination was sufficient to show that predictions cover the whole geographic space of the corpus area. It can be seen that the German-speaking area is well covered when it comes to individual predictions from trees. This means that single trees do, in fact, predict values close to the borders. This, in turn, means that different trees in the forest yield predictions on the other side of the geographic space, which finally leads to predictions in only a narrow part of Germany. The dots plotted in Fig. 3.9b possess an opacity of only 10%. This is done to keep the borders of the map visible and to recognize accumulations of predictions since some positions will be predicted more than once.

In general, a prediction can be calculated above the baseline for each direction. However, the improvement over the baseline in experiment 2 was low. It amounts to roughly



(a) The final RF predictions for each /z/ available in the corpus (blue points; 45,869 observations).

(b) Individual prediction of each tree in a forest plotted as a gray dot (with arbitrarily combined longitude and latitude coordinates from two models).

Figure 3.9: Visualization of the distribution for the prediction of the speaker positions based on the phoneme /z/.

10 km for longitude and roughly 26 km for latitude. How this reduction relates to the geography of Germany, can be seen in Fig. 3.10. The low improvement over the baseline is somewhat surprising, especially if one takes the good correlation between prediction and real values in the north-south direction into account.

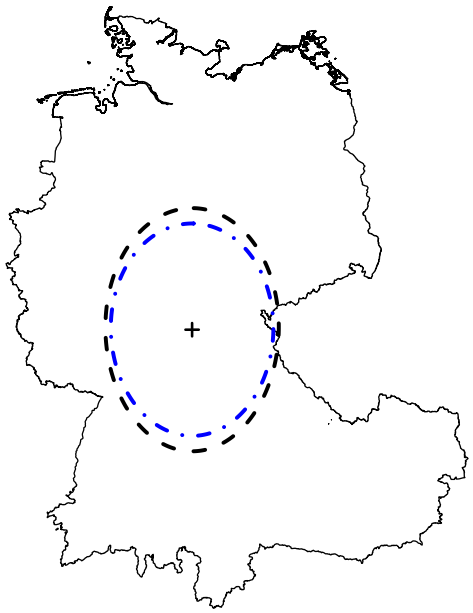


Figure 3.10: Midpoint of the GT corpus (black cross), the null model error (black dashed ellipse), and the error that resulted from predicting speaker positions with the RFs (blue dashed-dotted ellipse).

When compared to the classification, the regression uses similar information. This holds true for both phonemes and features and, for the north-south direction in particular, the selected features in phoneme /z/ are almost identical. The fact that the same phoneme performs best might be taken as evidence for the regression suffering from variation not happening in an axis-parallel fashion. Other explanations are that variation in the east-west cannot be modeled without considering the north-south distinction as well, or that variation in the east-west direction is not modeled sufficiently by the current feature set.

Fig. 3.11 illustrates the MAE in both a north-south and an east-west direction for all phonemes. The phonemes with a prediction above the baseline in both directions are plotted in black. Phonemes performing worse than the baseline in the east-west direction are plotted in orange and worse in the north-south direction in blue. It is interesting

that eight phonemes led to predictions that are worse than the baseline in the north-south direction, as in the North/South classification task all phonemes allowed a prediction above NIR.

Fig. 3.11 also shows a moderate correlation between the MAE in both directions (0.5014). This means that a phoneme that works well for longitude is likely to also work well for latitude and vice versa. As mentioned before, this might be an artifact of regional variation not being axis-parallel.

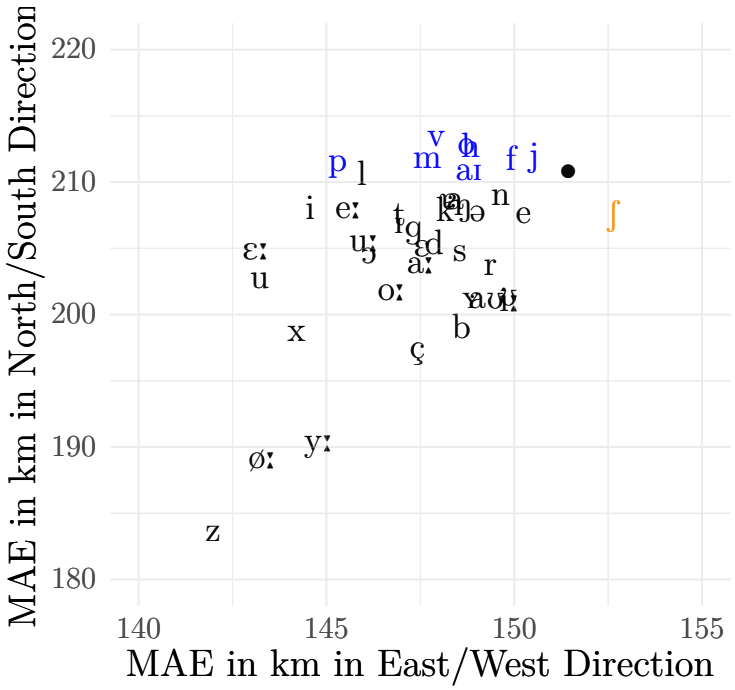


Figure 3.11: The phoneme-wise MAE plotted for both directions (east-west on the x-axis, north-south on the y-axis; lower values are better). The baseline using the null model is marked as a black circle. Phonemes that performed better than the baseline in both directions are black, phonemes worse in longitude are orange, and phonemes worse in latitude are blue. Please note, the x- and y-axes are scaled differently.

The bad improvement over the baseline is taken as evidence that the current method is not suitable to sufficiently model regional variation over the corpus area in a continuous fashion. The fact that more than one phoneme is necessary for an exact prediction would be in line with the semi-continuous changes of certain pronunciation variants, where many

changes occur across Germany in different linguistic variables and their pronunciation (representing different phonemes).

3.10 Experiment 3 – Combination of Features of Multiple Phonemes

3.10.1 Experimental Design

In the previous experiments it was shown that many phonemes can be used to predict 1) the North/South and East/West class above the level of chance and 2) a continuous position slightly above a chosen baseline along the longitude and latitude direction in the German-speaking area. These studies only used features of a single phoneme to model the regional variation over the corpus area. The following study will use much more information by combining all phonemes uttered by each speaker.

The data on different phonemes can be combined, for example, on a word- or a turn-level. Unfortunately, for both combinations, it is unlikely that 30+ different phonemes will be realized within them. Assuming that a level could be found on which all phonemes always occur, the second question arises, as to on how to combine multiple realizations. Combining the feature vectors of every uttered phoneme with any other uttered phoneme would lead to a tremendous amount of combinations. Moreover, adding the same feature vectors thousands of times in different combinations does not increase the information (i.e., the regional variation captured in the features vectors).

One solution circumventing both problems is to average out all realizations of a phoneme uttered by a certain speaker separately and then combine all these phoneme-feature vectors to form a single fixed length feature vector per phoneme. The following section will use this approach. For the material from the GT corpus, this results in the combination of 656 features of 33 phonemes. The resulting feature vector has a length of 21,648 and one vector is present for each of the 641 speakers in the corpus. The following 33 phonemes were combined: /ə, ɐ, a, aː, aɪ, b, ç, d, eː, ɛ, ɛː, f, g, h, iː, ɪ, j, k, l, m, n, ŋ, oː, ɔ, r, s, ʃ, t, uː, ʊ, v, x, z/. This procedure combines an average of 3751 phoneme realizations for each

speaker of these 33 phonemes.

This single fixed feature vector per speaker can be thought of as the "speaker-identifying" feature vector, as it combines all information on a certain speaker present in the corpus for all 33 phonemes uttered by every speaker. This identifying vector, therefore, contains information about the speakers individual characteristics, idiosyncrasies, sociophonetic variation, regional variation, etc. Using this information, the ML algorithm will then try to extract the information most relevant to model the variation along the respective geographic dimension.

The following experiment involves three steps:

1. Reducing the large initial feature set, based on the VI of a RF, to include only those features of those phonemes that are relevant for a prediction (for each direction)
2. Confirming that the resulting feature set for each direction a) is one that contains information about regional variation and, therefore, can be used to estimate a speaker's origin and b) is not an ML algorithm specific set, by evaluating the prediction accuracy of the subset by applying an independent learning algorithm (in this case an SVR).
3. Training a binary DT using the subset to see which features were relevant for which part of the prediction and, therefore, enabling an interpretation of the generated feature set. Additionally, the output of the DT is used to relate the features available to phonetic/dialectal phenomena in the German-speaking corpus area.

To some extent, this approach is inspired by Woehrling et al. (2009). In this study SVMs and DTs are compared regarding their performance. This study used a DT to relate the features to dialectal variation, despite the SVM yielding better results for more data.

3.10.2 RF – Results and Feature Selection

The previous experiments have shown that the choice of hyperparameters in the RFs only resulted in marginal differences. Therefore, no hyperparameter tuning is performed in this experiment. Instead, the parametrization most beneficial for the feature selection and the following explanation of the resulting feature set is chosen. In the current case, this

is a model that often reuses features that are (ever so slightly) better than others and, therefore, a model using only a limited set of different features.

The parameter influencing model complexity is *mtry*. For high values of *mtry*, more features are taken into account randomly at each split. Therefore, for higher *mtry* values the chance increases that a feature is selected multiple times in different trees and splits. Hence, high *mtry* values allow fewer features to be used in the RF, if there are features present that explain the regional variation better than others. Therefore, this parameter was set to $d/3$ (d is the dimensionality of the input feature set), which takes many features into account at each split and is the standard setting for regression.

To select the most important features and reduce the feature set under analysis considerably, an arbitrary cutoff point was defined as 1% of the maximal VI in the current experiment. That cutoff point was used to keep features in the subset if they have at least 1% of the best features' VI. Similar to previous experiments the RFs were validated using a Leave-25%-Speaker-out CV.

East-west direction: For longitude, the RF was able to predict the 641 speaker locations with an MAE of 120.47 km (1.6857°) and a strong correlation of $R = 0.6344$. The aforementioned feature selection method resulted in 408 different features, including features of 32 phonemes (out of 33). This means that only for the phoneme /iː/ no feature had a VI higher than 1% of the maximum VI.

It is noteworthy that of the 408 features used, 299 were Δ (153) and ΔΔ (146) features. The existence of many highly ranked Δ and ΔΔ features is surprising, as in previous experiments based on features of single phonemes, they were ranked lower.

North-south direction: For the latitude direction, the proposed method for feature selection resulted in 63 features from nine different phonemes. These were /ç/, /eː/, /ɛː/, /ɪ/, /n/, /r/, /s/, /v/, /z/. For this direction the MAE was 111.14 km (0.9992°) and the correlation was strong at $R = 0.8218$.

Of the 63 features used in this direction, 32 were Δ and ΔΔ features. These findings once again deviate from the findings for the top features in the single phoneme prediction

experiments.

Interim conclusion: As expected, averaging over multiple realizations of the same phonemes per speaker and combining the features of multiple phonemes, improves results compared to using the features of only one single phoneme.

Based on the arbitrary cutoff point of 1% of the maximal VI, a much smaller subset can be created. For longitude, 408 (1.8847%) of the original $d = 21{,}648$ features were kept in the reduced feature set, for latitude only 63 (0.2910%). It is interesting to see that in this experiment the Δ and $\Delta\Delta$ features are selected more often than in previous experiments. After averaging over multiple realizations, it seems that the Δ and $\Delta\Delta$ features can be used to describe regional variation. This is interesting insofar as the $\Delta\Delta$ features especially account for the transition parts of phonemes, which should not generalize well to different contexts due to coarticulation.

3.10.3 SVR – Results

Training SVR models using the resulting feature subsets should ensure that the sets describe regional variation in German speech, which can be exploited for speaker localization. This additionally will show whether the subsets are algorithm specific. The SVR used an RBF kernel and, untypically for SVR models, no hyperparameter search was performed. The tuning of hyperparameters seemed unnecessary as the performance of the SVR with the standard parameters already resulted in good performance. The achieved performance is sufficient to prove the selected features were a valid subset. Parameters used for training were $C = 1$ and $\gamma = 1/d$, where d is the number of features in the respective set (longitude: 408; latitude 64).

East-west direction: The SVR was able to predict the speaker locations with an MAE of 96.14 km (1.3452°) and a strong correlation of $R = 0.7613$. This is an improvement of 55.3 km over the baseline and 45.85 km over the best-performing single phoneme model.

North-south direction: For the north-south direction an MAE of only 96.94 km (0.8716°) and a strong correlation of $R = 0.8477$ was obtained. This corresponds to an improvement of 113.95 km over the baseline (cf. Sec. 3.9.2) and 87.26 km over the best-performing single phoneme model (cf. Sec. 3.9.4).

Interim conclusion: The SVR outperforms the RF in both directions, based on the features that were selected with the help of the VI. This is taken as proof that the feature selection indeed resulted in valid subsets containing sufficient evidence about the speakers' origin. This feature set will be used in Sec. 3.10.4, to train a DT which will hopefully shed some light on the relationships between the selected features. The resulting reduction in prediction error can be seen in Fig. 3.12.

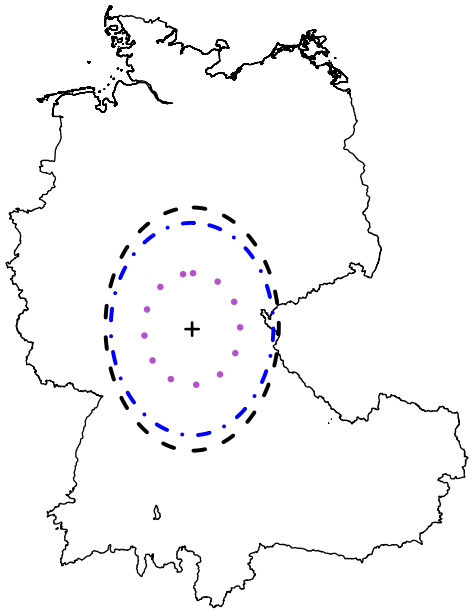


Figure 3.12: Midpoint of the GT corpus (black cross), the null model error (black dashed ellipse), the error of the regression for one single phoneme for the best phoneme /z/ with the RFs (blue dashed-dotted ellipse), and the error resulting from the SVR models based on the reduced, combined feature set (purple dotted ellipse).

3.10.4 Decision Tree – Results

Predictions based on the combined feature set lead to an improvement of results for both algorithms, i.e., the RF, using the full feature set, and the SVR, using the reduced feature set. Unfortunately, both these methods have drawbacks when it comes to the interpretation of the results.

Generally, the results of the RF are interpretable by the use of the VI, even if the VI has drawbacks as mentioned earlier. On the positive side, it should be noted again that the VI is a good indicator of important features. This was shown by selecting features based on the importance assessment of the VI and using the resulting feature subset to successfully predict speaker origins. Unfortunately, it only shows the overall importance of a feature, but not how it was used to divide up geographic space.

However, exactly the latter would be interesting for two reasons. First, to see how geographic space is actually divided by the features. That is: do features separate large areas that can directly be used for a rough estimation of the speaker origin, or is it necessary to put together the prediction based on small "islands". Second, to map those splits to already known dialect boundaries and variation. To overcome this particular shortcoming of SVR and RF, a DT is trained using the reduced feature set obtained in Sec. 3.10.2. The splits in the generated tree enable a direct interpretation of features within the geographic space that they occupy.

As stated in Sec. 3.7, a binary DT was used, and one model was trained for each direction. As with the RFs and the SVR models, the DT was trained to predict the longitude and latitude of the speakers' origins. The results of the two models are:

East-west direction: For longitude, an MAE of 142.06 km (1.9878°) and a correlation of $R = 0.3886$ was achieved with the DT. This is much worse than the prediction based on the same dataset with RF and SVR.

North-south direction: For latitude, the prediction results in an MAE of 126.39 km (1.1363°) and a strong correlation of $R = 0.7317$. The resulting DT generated is shown in Fig. 3.13.

Interim discussion: The bad performance of the model for the east-west direction might stem from the fact that for this direction the differences are not the same for the North as for the South half of Germany. A DT would be unable to model this efficiently, as features that do not reduce impurity on their own, will not be used for a split at a high level. This does not mean it is impossible that another feature is selected high in the tree's hierarchy, for example, due to a lack of better features, and is then later refined by such a feature. However, this will not always happen. The SVR using an RBF kernel, on the other hand, would be less influenced by this, as the decision boundary could be modeled to possess arbitrary shapes in any subspace.

To verify the hypothesis that it is easier to predict the east-west direction separately in the North and in the South half of the corpus area, two separate models for each half were trained for all three algorithms (RF, SVR, and DT). The results of those models can be found in App. A.4. In summary, the correlation rises to $R = 0.4950$ and $R = 0.5333$ and the MAE decreases to 123.02 km (1.7214°) and 127.22 km (1.7801°) in the North half and South half respectively. This is taken as evidence for the validity of this hypothesis.

Due to the bad performance of the unified longitude model, it is unclear whether the generated DT sufficiently models the existing variation in the east-west direction. Therefore, a further analysis of the resulting model is omitted.

For the north-south direction, the prediction works better than for the east-west direction. The top feature used for the initial split VC (0) can be found among the top features in all three experiments in this chapter (cf. Fig. 3.13). It, therefore, seems to be a stable feature across experiments and settings.

3.10.5 Phonetic Interpretation of the Decision Trees

General Notice: The following section attempts to phonetically interpret some of the features that have been selected during the training of the DT. This is done based on the properties of the features and previous studies on regional variation in Germany. An auditory, manual check on whether the interpretation can be validated by the inspection of the visualized signal and based on human perception was performed only on small, random

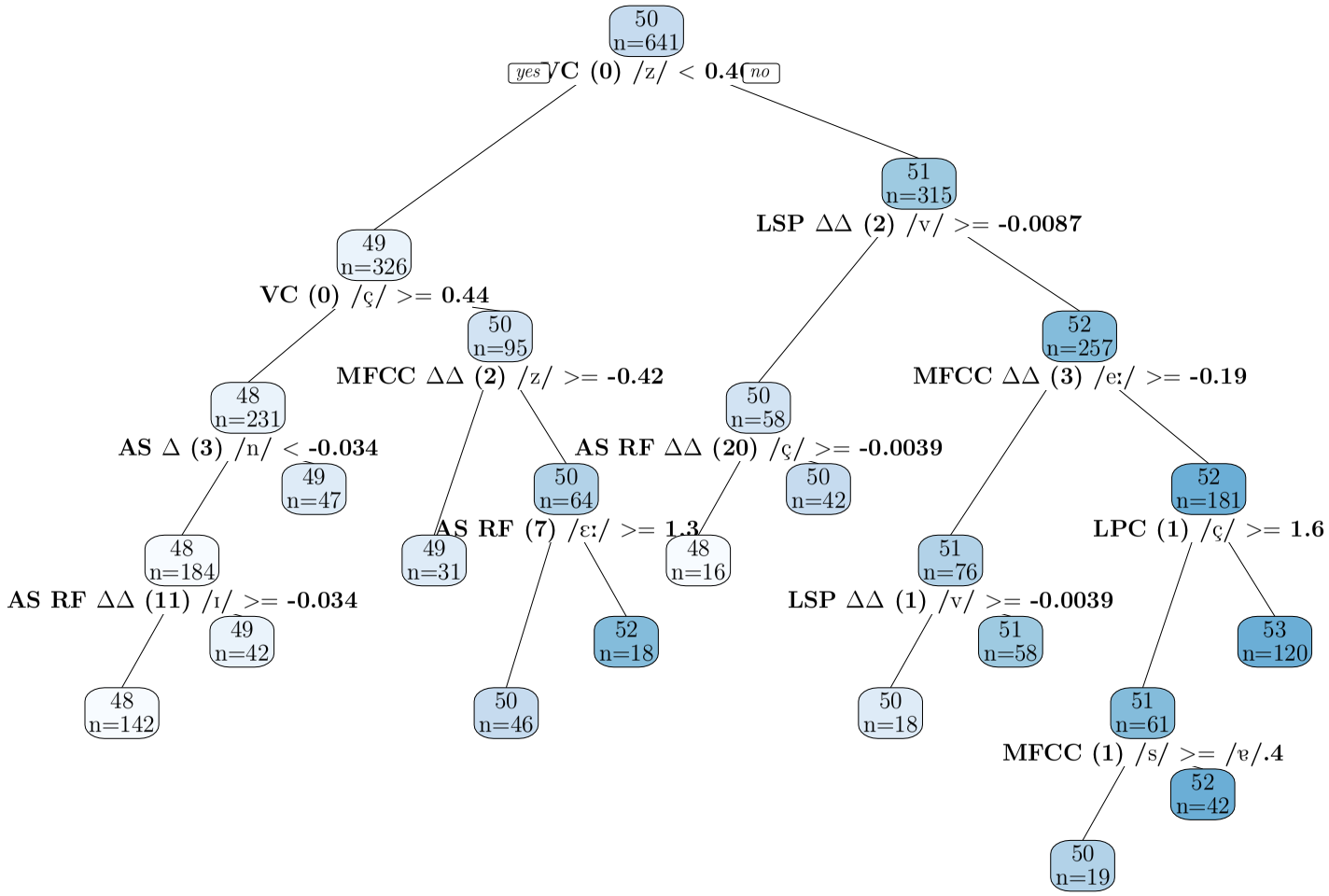


Figure 3.13: Decision tree for the north-south direction in Germany. The color of nodes and leaves correspond to the output variable "latitude". Brighter colors mean lower values (South), darker colors higher values (North). Values used for splitting are rounded to two decimals for better readability (for the original values cf. App. A.3).

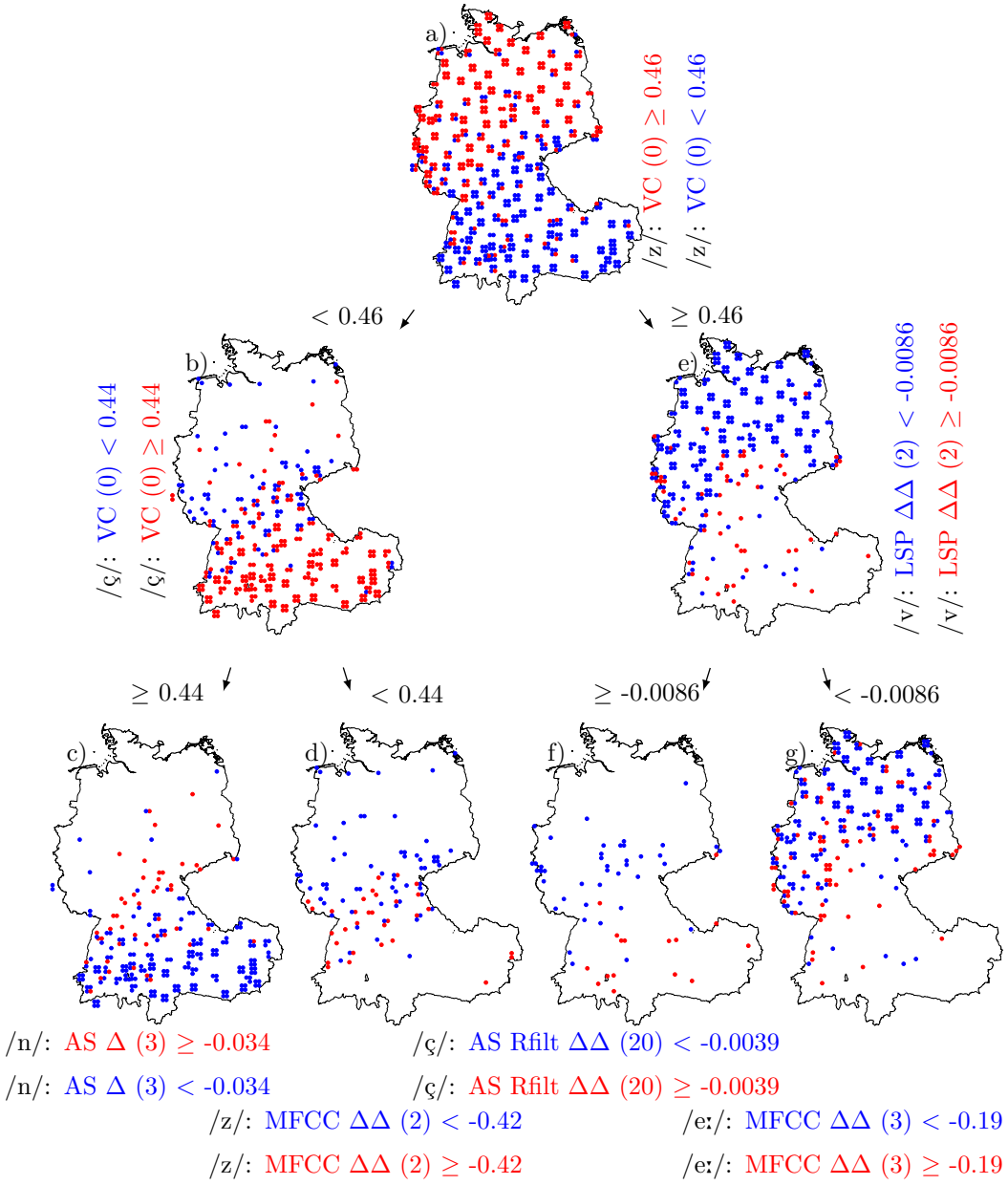


Figure 3.14: Visualization of how the geographic space is divided based on a DT. The split variables and values can be found next to the according map. Values below the threshold are blue, values above the threshold are red. Values used for splitting are rounded to two decimal places for better readability (for the original values cf. App. A.3). For each split, the feature name is shown. The letters a) to g) are used for reference.

subsamples of the dataset. Taking the amount of data into account, it would be unfeasible for the current study to manually confirm the suspected phenomena across all speakers systematically.

/z/ VC (0): The initial split in the north-south direction is done using the VC (0) in phoneme /z/ as seen in Figs. 3.13 and 3.14. This feature alone already results in a good division of a northern and southern part of the corpus area, which is likely due to the devoicing of /z/ in southern varieties (cf. Sec. 3.8.4 and boxplots of feature values in Fig. A.17). How this feature divides the geographic space can be seen in Fig. 3.14a (for a more detailed visualization of the same phenomenon cf. Fig. 3.4).

/ç/ VC (0): The southern speakers are then further split based on the VC (0) of phoneme /ç/, which is visualized in Fig. 3.14b, which shows that this feature results in a good division between speakers originating from the Middle/North of the corpus area, as well as those speakers that have a low voicing probability and originate from the South (cf. boxplots of feature values in Fig. A.18).

Barbour et al. (1990, p. 154) report that /ç/ is produced in a more velar fashion in the South (as, e.g., in <durch> which is pronounced in its canonic form in the North like /dʊɐç/, whereas it might be pronounced as [dʊɐx] in the South), leading to a [x]-like pronunciation. When comparing the realizations of the phonemes /ç/ und /x/ for all speakers, the feature VC (0) seems to be higher for /x/[18] (cf. boxplots of the values for both phonemes in Fig. A.19). This would, therefore, agree with the southern informants producing a more velar sound.

Additionally, on randomly inspecting a few examples of /ç/ above and below the split point, it appears that two factors add to this effect. First, the elision of /ç/, which results in a wrong segmentation and, in turn, measuring the voicing of an adjacent vowel instead of /ç/ (e.g., /ɪç/ → /ɪ/). As a reminder, the automatic S&L was performed in forced-

[18]Ignoring the fact that the phoneme class /ç/ might also contain [x]-like realizations. It, once again, seems worth noting that the features extracted from the speech signal might behave differently than expected. For example the phoneme /x/ is not considered to be voiced. Nevertheless, it shows a higher VC (0) than /ç/. As long as this effect is systematic, this does not pose a problem in the current study.

alignment mode, i.e., no adaption of the canonic form as output by G2P was performed. Second, the realization of [k] and [kç] instead of [ç], where the aspiration of the plosive produces a rather high value in VC (0). This is somewhat surprising as aspiration is not voiced. An example of this is the word-initial Standard German /ç/, as in <Chemie>, which might be pronounced as [k] or [kç] in the South (e.g., Brinckmann et al., 2008; König, 1989, p. 97-98).

/n/ AS Δ (3): The speakers in the subset with a VC (0) higher than 0.44 (extracted in a phoneme /ç/), are split based on the AS Δ (3) (221.76 Hz − 411.84 Hz) of phoneme /n/ as seen in Fig. 3.14c. This is slightly above the frequency range for which the F1 of nasals is expected (roughly at 200 Hz, Machelett, 1996). The split results in a subset of speakers that are predicted to originate from the middle of the corpus area and a southern part that will be split further. The Δ of the AS (3) is a bit less steep in the negative direction for speakers from the middle of the corpus area (even though never equal to 0 or positive; cf. boxplots of feature values in Fig. A.20). That means that the decrease of energy in this band is smaller for speakers from the middle of the corpus area, than for those from the South. At the time of writing, no phonetic interpretation could be found that would explain this phenomenon.

/v/ LSP ΔΔ (2): On the side of the tree seen in Fig. 3.13, where speakers possess a high voicing probability on the phoneme /z/, the speakers are divided by the curvature (ΔΔ) of the LSP (2) of phoneme /v/, which is visualized in Fig. 3.14e. For speakers that originate further North, the frequency range described by LSP (2) is between 1709.15 Hz and 2460.42 Hz and for the southern speakers the frequency is between 1680.40 Hz and 2309.77 Hz. This is the region that lies around or above the frequencies for which a second formant in /v/ would be expected. The short-time functional ΔΔ might describe a change in the energy present in this part of the spectrum. Evidence for this can be found in the correlation of the feature with the log energy ΔΔ at $R = -0.63527$ and with SV ΔΔ at $R = 0.6671$ (where it is assumed that the SV ΔΔ is smaller if the energy stays constant in the current and neighboring frames).

This might be linked to a more or less voiced /v/. König (1989, p. 91) reports that northern speakers, in general, produce a clearer labiodental /v/ than southern speakers. Furthermore, he notes that his observations suggest that the more labiodentally the phoneme is realized, the more voiced it is. In this case, however, the energy present in the phonemes would suggest instead that speakers belonging to the southern group (red/blue in Fig. 3.14e) produce even more voiced consonants than the northern speakers. This might be due to the fact that this speaker group behaves differently than the literature would suggest, i.e. produce a more voiced /v/. This assumption is based on the fact that the speakers shown in Fig. 3.14e do not devoice their /z/ either. It might be possible that those speakers use more voicing in general.

This would agree with the feature LSP ΔΔ (2) and the reconstruction of the spectrum in a higher frequency range for northern speakers (as the reconstruction starts at higher frequencies due to a missing voice bar). This means that for the northern speakers, the reconstruction at higher frequencies would lead to higher values for LSP (2), which would explain the downward open parabola that is described by the ΔΔ feature. König (1989, p. 91) reports that not all varieties in the North produce the /v/s voiced 100% of the time. Some of the locations described do not even produce voiced /v/s 20% of the time. On the other hand, the ΔΔ of LSP (2) would stay more constant for voiced /v/, leading to smaller ΔΔ values (for the southern speakers), as the second formant appears around the same range a formant of vowels would be expected.

/ç/ AS Rfilt ΔΔ (20): The speakers that are placed in the group located in the Middle/South of the corpus area based on the split of the feature LSP ΔΔ (2) are split using the feature AS Rfilt (20) ΔΔ of phoneme /ç/, as seen in Fig. 3.14f. This feature describes the curvature at frequencies between 3805.03 Hz − 4734.02 Hz. At this frequency range, more energy would be expected in [ç]-like or [ʃ]-like realizations due to frication, compared to phones like [k]. Again the ΔΔ feature possibly describes a more or less voiced phone (the correlation between the "normal" voicing probability and AS Rfilt (20) is strong at $R = 0.60$). This agrees with using voicing probability in the same phoneme, even though none of the above segments are supposed to be voiced.

Additionally, the duration of the phones uttered by the southern speakers is shorter (between the features AS Rfilt ΔΔ (20) and duration a moderate negative correlation of $R = -0.53$ exists), meaning the more downward curved a value, the shorter it is. The shorter a phoneme is, the fewer feature vectors are averaged over the 20% midpoint. This means that if only one feature vector is used for the midpoint, small and irregular changes can stay in the final feature vector. For longer phonemes this is less likely, as more vectors are averaged and the chance is higher that local changes are canceled out by the following change in the opposite direction (cf. boxplots of feature values in Fig. A.23).

The feature AS Rfilt ΔΔ (20) could combine the two different phenomena, voicing and duration, into one feature, which better explains the target than the two separate features voicing probability and duration alone.

/eː/ MFCC ΔΔ (3): The speakers that are placed into the northern subset are split at the MFCC (3) ΔΔ feature in phoneme /eː/ which can be seen in Fig. 3.14g. The speakers in the North of the corpus area again have a more negative curvature than the group in the middle of the corpus area. In this case, the feature used for splitting does not correlate well with either duration (0.035) or voicing probability (0.18). The latter is especially supposed to be equally strong for both groups.

It can be seen that the speakers originating from the North part of the corpus area have a lower F1 (median 371 Hz) and a higher F2 (median 1780 Hz) than the speakers belonging to the subset located more in the middle of the corpus area (F1 median: 393 Hz; F2 median: 1749 Hz). This means that the latter group produces a more open /eː/ than the northern speakers (a more /ɛː/-like sound). On random inspection of some examples, it can be seen that in many examples of northern /eː/, as would be expected, less energy is present in the area between the first and second formant. However, it is unclear why the ΔΔ feature is chosen at this point and not the base feature (MFCC ΔΔ (3); cf. boxplots of feature values in Fig. A.24).

3.10.6 Prediction Error in Both Dimensions

In the following, the prediction error per recording site and per modeled dimension (longitude/latitude) is examined more closely. This is done based on the SVR as it resulted in the lowest MAE. Fig. 3.15 shows a) the individual predictions for all speakers, b) the prediction error per location, and c) the standard deviation of the prediction error at a location.

Individual predictions are shown in small pink dots on the map in Fig. 3.15. In the outer corners of the corpus area no such dots appear, as one would expect. In particular, in the South-East, the North, and the North-East, large areas emerge where almost no predictions were made.

In Figs. 3.15a and 3.15b each location is plotted with a circle, where both the size and the color indicate the size of the prediction error. The prediction error is the mean of the prediction errors of all speakers from the respective recording site. A larger size/a darker color indicates a larger error. The site label is printed in cases in which the error is higher than 30% of the maximum error for all sites.

In order to distinguish between large errors that result from an inhomogeneous group in a location, and large errors over all speakers in a location, the standard deviation is shown in Figs. 3.15c and 3.15d. As in Figs. 3.15a and 3.15b, size and color is an indicator of the size of the standard deviation. Labels are printed if the standard deviation exceeds 1.

As is to be expected, especially in areas where no individual predictions are made, large errors occur. This is the case near the minimal and maximal values of the prediction intervals. In those areas, wrong prediction errors lead to large errors and are, therefore, avoided by the algorithm.

However, it can be seen that large errors also occur in the middle of the map. An error in such a location is more likely to belong to a speaker or a set of speakers that were hard to recognize based on differing regional variation. For big errors in the middle of the corpus area, two possibilities exist.

First, in some locations, the errors in the site are quite large, and the standard devi-

(a) Predction error in east-west direction.

(b) Predction error in north-south direction.

(c) Standard deviation of above error in east-west direction.

(d) Standard deviation of above error in north-south direction.

Figure 3.15: Visualization of the prediction error and standard deviation of this error over all locations of the corpus area, based on the SVR model trained on the combined feature set. The size of the circles is proportional to the prediction error in a) and b) and proportional to the standard deviation in c) and d). Labels are printed if the prediction error is larger than 30% of the maximal error in a) and b), and if the standard deviation is larger than 1 in c) and d). Size and color of each point indicate the magnitude of the respective measure, white/light colors indicate low values and dark blue colors high values. Pink dots are individual predictions.

ation is comparatively small. A small standard deviation indicates that all speakers were predicted with roughly the same accuracy. In these locations the prediction was difficult, likely due to regional variation that was not modeled correctly or variation that deviates from those of the surrounding sites.

Second, there exist sites in which inhomogeneous predictions were made across individual speakers. Those sites possess a large prediction error and a large standard deviation between speakers. Hence, one or several speakers were more difficult or easier to localize than others in the same location.

An example of a speaker location that was hard to predict for all speakers in the east-west direction was "NDH" (Nordhausen). At this site, the average error was comparatively large, as it is located in the middle of Germany, but the standard deviation is low. This means that all speakers were recognized equally poorly. An example in the north-south direction is "ZIT" (Zittau) in the very east of Germany. "ZIT" is located in quite a central position when it comes to latitude, but still exhibits a large error, while the standard deviation is small.

3.10.7 Discussion of Experiment 3

For all three applied machine algorithms, good prediction results could be achieved by combining the available features for multiple phonemes. For the DT however, this only holds true for the north-south direction. This and the previous results, in which the relative improvement of the east-west always falls behind the north-south direction, confirms the assumption that the east-west direction is harder to predict than the north-south.

This finding is reflected in traditional dialectology, in which the hierarchical grouping of dialects is initially performed in the north-south direction into two large groups: Low German (*Niederdeutsch*) and High German (*Hochdeutsch*) based on the Benrath Line, which describes whether the High German sound shift took place or not (Barbour et al., 1990, pp. 33–35 and 76). Often, the High German area is split up further into the groups: Central German (*Mitteldeutsch*), and Upper German (*Oberdeutsch*), based on the *mitteldeutsch/hochdeutsche Sprachscheide* (MHS; free English translation: Middle Ger-

man/High German Language Boundary; Lameli, 2008a[19]). These large groups are then further divided into various levels of granularity, mostly in the east-west direction. Barbour et al. (1990, p. 85) also mention that the isoglosses separating East from West are not as clear-cut as those between North and South. An explanation for this might be that along the north-south axis not only dialectal, but also political and economic variations occur (Barbour et al., 1990, p. 81). This is even more true for dialects that span certain countries (e.g., Bavarian which is spoken in Germany and Austria, Auer et al., 1996, p. 15).

Taking this into account, the SVR was still able to make a good prediction in the east-west direction. This is probably due to the SVR being able to benefit from structuring features and the use of an RBF kernel that allows it to model arbitrarily shaped regression functions. An example of this kind of a structuring feature would be speaker sex. This feature is not directly usable for DTs, as it does not directly explain the target variable (therefore the decrease in impurity is small). This unfortunately holds true even if it, for example, could be used to differentiate between F0s of men and women. But, if the original feature is not suitable for predicting the target sufficiently on its own, it will not be selected early in the splits. Further down the tree, where it could be successfully used for splitting variables, the structure of the problem is a different one (as splits are carried out based on residual impurity). The SVR, on the other hand, can take information like this directly into account in its decision boundary (dimensionality expansion).

The $\Delta\Delta$ features that are used in this experiment seem to correlate well with two different effects. First, the duration of the phoneme. By averaging fewer feature vectors in shorter phonemes, small changes that are only taking place at a local level are less likely to be canceled out than in longer phonemes. Second, the difference between a voiced vs. an unvoiced realization of a consonantal phoneme. In phonemes that are realized unvoiced, the change in curvature is a different one, when compared to phonemes that are realized

[19]The MHS can also be seen in Wiesinger (1983) as the division between the Middle German dialects – Franconian, Hessian, Thuringian, and Saxonian – from the High German Dialects – Alemannic, East Franconian, and Bavarian. Barbour et al. (1990, p. 79) calls it the Germersheim Line. In the following MHS will be used.

voiced. This is based on the energy distribution in the spectrum (e.g., the lower parts where a voice bar exists for voiced consonants).

A DT was generated so that features could be related to the geographic space of the German-speaking corpus area. To visualize this in an intuitive way, the splits into speaker subsets were plotted on different maps. In these visualizations, it can be seen that acoustic features possess a geographic distribution that is similar to linguistic variables in traditional dialectology (e.g., Wrede et al., 1927–1956; Wiesinger, 1983; König, 1989). However, the boundary for the contrast of voiced vs. devoiced /z/ between the North and the South of the corpus area is located further south than reported in the literature by König (1989).

Nevertheless, the splits result in a geographic division of speakers that resembles a traditional isogloss. In Fig. 3.16 it can be seen that the north-south boundary from the current study, coincides with the MHS (for a definition cf. 3.10.7), i.e., the border between Upper and Middle German. This separation can also be found when looking at the split for phoneme /v/ in Fig. 3.14. It should, however, be noted that the DT bases its splits on the maximal decrease in impurity. The splitting point in the feature VC (0) could, therefore, be placed at a feature value that separates the corpus area well in two equal parts. Dialectologically, this splitting point is somewhat arbitrary. Nevertheless, it is interesting for two reasons: firstly, it is taken as evidence that the proposed method (only relying on acoustic features) is capable of modeling dialectal variation and, secondly, it can be seen that young speakers from the first decade of 2000 behave similarly (at least to some extent) to those in previously reported studies.

3.11 Discussion of Speaker Origin Estimation

It has been shown that RFs can be used to roughly assign speakers to the two halves of the corpus area, from which they originate. Localization works better for North/South classes (70.37% accuracy for the best-performing phoneme /z/) than it does for East/West (57.91% for the best-performing phoneme /øː/).

The classification approach from experiment 1 performs poorly when compared to the results of previous studies (cf. Sec. 3.3). This has several reasons and will be discussed

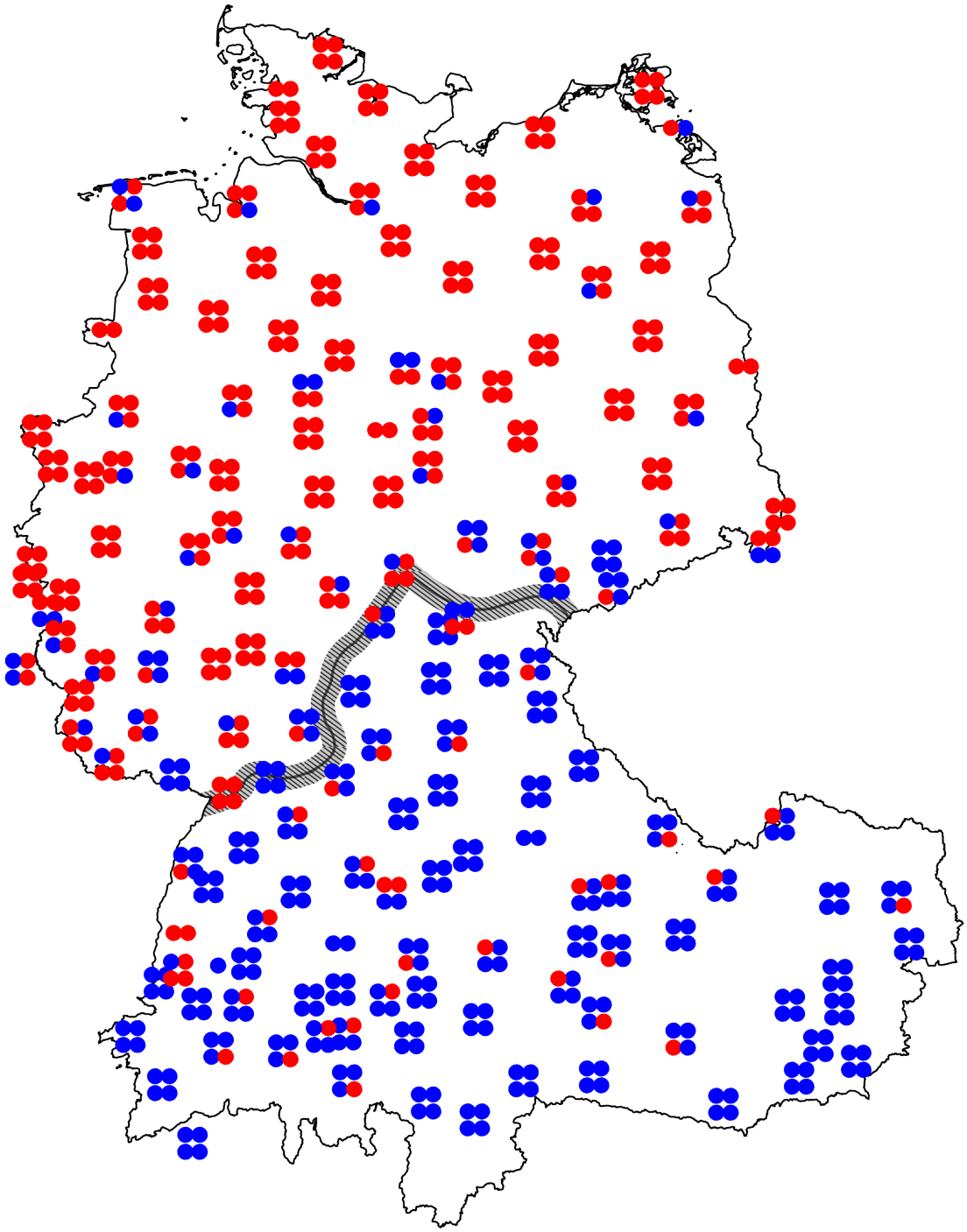


Figure 3.16: Overlay of a rough approximation of the MHS (according to Lameli, 2008a) and the initial split of the corpus area in the current study based on the DT (cf. Fig. 3.14a).

more thoroughly in the next few paragraphs, as the previously mentioned methods are not directly comparable.

First, only the methods based on spontaneous speech can be compared, as studies have shown that spontaneous speech is harder to classify than read material (e.g., Brown, 2015). The studies using read speech have the advantage of clearer pronunciation, which in turn makes the processing and detection of phenomena easier.

Second, no further preselection of words, contexts, or POS was performed on the phonemes before feature extraction. This is likely to result in more noise in the data than a careful selection of target words. This was done in the current study to adhere to the bottom-up approach.

Third, the previous studies presented in Sec. 3.3 all use material with distinct dialect or accent labels. Speech that is supposed to represent a certain dialect region is likely to be recorded at the center of a certain dialect area in which the desired dialect is spoken (e.g., in the ABI corpus; D'Arcy et al., 2004), and informants are selected that *represent* that dialect (although this is a subjective categorization made by the corpus creators). This, by definition, results in larger differences between speakers, than having recording sites spread out well over the recording area (which are located in close proximity to each other).

That being said, special corpora exist that try to model the variation in a small region and do not possess the just mentioned deficiency. An example for this kind of corpus is the AISEB corpus (Watt et al., 2014) that captures variation at the English/Scottish border. The AISEB corpus possesses good local resolution, but unfortunately only covers a small region. In the GT corpus the recordings sites possess both, good resolution over the corpus area and coverage of a large area. Furthermore, the subjects only had to fulfill a few criteria (their education, the origin of their parents), but other than that no preselection was performed based on, for example, how strong their vernacular is.

Fourth, in the current study shorter speech samples were used when compared to previous studies, to make the distinction between the dialect labels in experiment 1 and to predict a speaker's origin in the regression case in experiment 2. In the first two experiments, the time span of a sample used to make a prediction (being equal to a fraction of

one phoneme) was several orders of magnitude smaller than previous approaches (roughly 30 s − 45 s for text-independent methods and 3 m − 13 m for text-dependent approaches; cf. Sec. 3.3).

And finally, using the method in this chapter permits explaining phonetic/dialectal phenomena underlying the predictions. This is not possible with most other approaches. Comparable approaches to experiment 1 are, for example, Brown (2015) obtaining 52.5% accuracy on a 4-way classification task and Woehrling et al. (2009) achieving 85% in a 3-way classification task. Relative to these results, the current performance moves to a more acceptable range, at least for North/South classification.

Based on the promising results in the classification experiment, the reduction achieved in experiment 2 over a conservative baseline using RFs is disillusioning. The improvement equals 9.45 km (6.24%) for longitude and 26.69 km (12.65%) for latitude. For both, the classification and the regression task, similar features appear within the top ten features. As the tasks are inherently different, this is somewhat surprising. However, as the semi-continuous changes of certain pronunciation variants that do not all occur in the same phoneme, it is not surprising that single phonemes are not suitable to correctly estimate a speaker's origin. The small decrease in error compared to the baseline make it unlikely that this method can be successfully employed to improve ASR performance.

One way to improve localization performance could be the use of *dynamic* instead of *static* features. In many phonemes, for example stops, the change of the speech sound over the course of their duration could be captured more adequately using dynamic features (Reichel, personal communication, 2012–2018). This could, for example, be achieved by taking the first n coefficients of a DCT that is applied to the features throughout the duration of a phoneme. One drawback of this would be the drastic increase of the amount of features by n (for each feature, n new features would be created).

When it comes to the pure performance in estimating the speaker origins, "black-box" approaches could also be considered. These methods come with better internal modeling capabilities at the cost of explaining variation. Examples of these are, for example, Deep Neural Networks (DNNs) (Lopez-Moreno et al., 2014) or i-vector approaches (DeMarco et al., 2013). Lopez-Moreno et al. (2014) report that 10 h of speech are necessary for their

DNN to outperform an i-vector approach. The map task recordings of the GT corpus comprise roughly 67 h (cf. Sec. 3.5.1), which, therefore, could be sufficient to train a DNN.

Another aspect that influences the performance is intra-speaker variability. The fact that speakers are unable to reproduce a speech sound in exactly the same way, contributes to wrong classifications and estimations. However, that informants from the same or proximate regions do not always behave equally, and that informants themselves do not always produce variation to the same degree, is a problem that all studies on regional variation face (e.g., cf. the differences of pronunciation in many regions in the maps of König, 1989).

Experiment 3 evaluated how the combination of the data from multiple phonemes changes the results. It was found that localization based on a large feature set leads to a considerable improvement in the results. The best performance could be achieved using SVR on the reduced feature set with an MAE of 96.14 km in the east-west direction and 96.94 km in the north-south direction. However, in a real-world application (e.g., voice commands to control a smartphone, dictation of e-mails, etc.) it is unlikely that the required amount of speech material for this approach is available.

The feature selection performed in experiment 3 resulted in a subset containing sufficient information to allow for speaker localization. The SVR model based on this set even outperformed the RF using the full feature set. This can be attributed to the method itself, as it is able to model more complex regression functions. The selection itself, by retaining the features that have at least 1% of the maximal VI output by the RF, resulted, therefore, in a valid subset. However, for the two directions (north-south and east-west) it resulted in different sized subsets. It is interesting to see that the initial feature set of 21,648 was reduced drastically to only 408 features for longitude (1.8847% of features retained) and only 63 for latitude (only 0.2910% of features retained). This can be taken as further evidence that the east-west direction is more complicated to predict.

The drop in VI in the east-west direction is less steep than in the north-south direction (cf. Fig. A.10). This, and the generally worse results of the prediction, can be taken as evidence that the east-west distinction is harder to predict as more features have to be combined to predict this direction. This is in agreement with traditional dialectology, as

dialects are categorized hierarchically first in the north-south direction into two groups (Low German and High German) or three groups (Low German, Central German, and High German), which are then split up further in the east-west direction (e.g., Wiesinger, 1983; Barbour et al., 1990, p. 79).

In experiments 1 and 2, in which only single phonemes were analyzed, not many Δ and $\Delta\Delta$ features appear in the top ten ranked features; a trend that is continued in the top 50 features as well. However, in experiment 3 many Δ and $\Delta\Delta$ features were ranked highly. Possible reasons are: a) the Δ and $\Delta\Delta$ capture more dynamics and are generally favorable, but are more influenced by phoneme context. This is a problem in case in which single phonemes are used for prediction, but this problem can be eliminated when these phonemes are averaged over multiple realizations. And b) the structure of splits is completely different, as the tree can choose from different features of more phonemes compared to previous experiments.

Based on the feature subset, a DT was generated so that features could be related to the geographic space of the German-speaking corpus area. The resulting splits were visualized on a map showing that the features indeed possess a geographic distribution similar to linguistic variables in traditional dialectology. Successfully employing a feature set that contains standard features (e.g. formants), but also non-standard and technology-driven features (such as MFCCs or PLPs), has shown that many of these features can be used to capture regional variation. A further advantage of these features is the robustness with which they can be extracted from a speech signal. However, this comes with a penalty regarding how complicated it is to relate these features to phonetic phenomena.

The shape of the geographic regions based on the splits in the DT, raises the question of how well clustering algorithms would perform. This was already outlined in Kisler et al. (2014). Regarding the splits based on a small number of features, clustering algorithms like k-means could already result in interesting patterns. Nevertheless, the high-dimensional clustering techniques mentioned in Kisler et al. (2014) would also be a good basis for an interesting study.

The splits resulted not only in regions where certain phonetic features are prevalent, but also resemble the geographic division of speakers into traditional dialectological regions.

Fig. 3.16 shows that the traditional dialect boundary dividing the Upper and the Central German dialects coincides with a division of speakers based on a DT split. At first sight, this might indicate a change compared to König (1989, p. 93–96), as in his examination the change between voiced and unvoiced variants occurs further North. He describes recording sites that pronounce /z/ with "strong voicing 10%" and "weak voicing 10%" of the time. That being said, the value used for the split in the DT is somewhat arbitrary and is chosen to best separate the North from the South of the corpus area, to minimize error during training. It is not known how the informants in the GT corpus would be categorized according to this classification, missing the manual auditive validation of the voicing in /z/. Indeed, voicing has a strong north-south variation, and young speakers from the early 2000s seem to use this feature, which is in line with the behavior reported in the literature. The close resemblance of distributions from traditional dialectology, mostly working on written transcripts of informants' speech, and the strict bottom-up approach pursued in the current study, based only on acoustic signals, is taken as validation for both approaches with respect to each other.

Chapter 4

Measure of Confidence for Corpus Analysis (MOCCA)

This chapter builds on the following previously published work:

Thomas Kisler and Florian Schiel (2018a). "MOCCA: Measure of Confidence for Corpus Analysis - Automatic Reliability Check of Transcript and Automatic Segmentation". In: *Proc. LREC*. ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA)

4.1 Abstract

Automatic Speech Recognition (ASR) systems estimate the quality of their recognition results using methods called Confidence Measures (CMs). The present study examines the applicability of these CMs for automatic corpus analyses on the transcription and segmentation level. In order to do so, Machine Learning (ML) algorithms and features that have been proven to work in ASR CM are examined with regard to their performance in predicting two quality measures: the correctness of the transcript and the quality of the

subsequent alignment. It is shown that the used methods are applicable to the automatic evaluation of transcripts of speech corpora, with an accuracy of 78% for detecting words labelled incorrectly, and, to a lesser extent, to assess the quality of an automatic MAUS segmentation and labeling (S&L) with a correlation coefficient of $R = 0.60$. The proposed method introduces an S&L post-processing step based on Support Vector Machines (SVMs) for both tasks, whereas the applicability of Random Forests (RFs) is also studied. Different parametrizations for both algorithms are analyzed based on 10-fold Cross Validation (CV) using a corpus of spontaneous speech in German and are evaluated against a second, independent corpus.

4.2 Introduction and Motivation

The creation of a new speech corpus typically involves three major steps: (1) the recording of a speech signal, (2) the orthographic transcription of the recordings, and (3) the alignment of a phonetic transcription to the recorded signal. Step (3) is referred to as segmentation and labeling (S&L). The quality of these three pre-processing steps has a big impact on the usefulness of the resulting speech resource. Step (2), the transcription of a speech recording, can be done either manually or via Automatic Speech Recognition (ASR). In both cases, the transcription will most likely contain errors in the form of deviations between the transcribed and the actually spoken words. Step (3), the S&L, can also be done either manually or automatically, based on the transcript created in step (2). Manual S&L is even more time consuming than the manual transcription process. Based on experience, it is around 20 to 100 times slower (Draxler et al., personal communication, 2016). This step is often done in two parts: initially, an automatic alignment, for example, using a forced-alignment or a similar technique, is executed and the boundaries and labels of this automatic alignment are then manually corrected afterward. In both cases, the manual correction of the transcript and the manual correction of the S&L, the auditory check of all recorded speech signals by a human transcriber is an expensive, time-consuming, and error-prone task. This leads to corpora often not being checked and corrected by a second human expert.

Therefore, quality measures that automatically detect parts in which potentially 'something went wrong' are of interest for tasks 2) and 3). This kind of automatic method could reduce time and effort considerably by automatically detecting erroneous parts, as the manual correction process would benefit greatly from an automatic way to find erroneous segments.

The two problems examined in this study are strongly related to estimating the correctness of a hypothesized word sequence $\hat{W}$ in an ASR system. Measures that can make a statement about the correctness of recognition hypotheses are called "Measure of Confidence", "Confidence Measure", or "Confidence Estimation" (Seigel, 2013; Jiang, 2005). In the following, the term Confidence Measure (CM) will be used. CMs can either be produced during or after the recognition stage, depending on how the CM is implemented (i.e., integrated into the decoding process or in a post-processing step). Research on confidence measures for ASR has attracted significant attention in the past (e.g., Schaaf et al., 1997; Weintraub et al., 1997; Kemp et al., 1997; Zavareh et al., 2013; Jiang, 2005; Pellegrini et al., 2010; Parada et al., 2010; Chen et al., 2013; Seigel, 2013; Ghannay et al., 2015). CMs are mainly used to detect recognition errors and to my knowledge, only a single study has been conducted so far that applies methods used in ASR CM to assess the quality of an automatic S&L (Paulo et al., 2004), which will be discussed in Sec. 4.3.6.

This leads to the following general research question for the current study: *Can methods that have been successfully applied to estimate the quality of the recognition hypothesis of ASR systems and the quality of phoneme-level S&L be used for an automatic quality assessment of the transcription of speech and automatic S&L on word-level*?

The remainder of this chapter is organized as follows: the next section provides an overview of work done in the field of confidence measures and some relevant work for the present study is discussed in depth. Sec. 4.4 explains in detail how the CM estimation was implemented in the current study (which features were used, which ML algorithms were applied, and which data was used for training and testing). In Sec. 4.5 the experiments are explained and their results discussed. Finally, Sec. 4.6 discusses and concludes the findings of the current study and Sec. 4.7 gives some directions for future research and system improvements.

4.3 Confidence Measures

4.3.1 Introduction to Confidence Measures

As stated before, CMs try to estimate the quality of an ASR output, namely the hypothesized word sequence $\hat{W}$. Two different types of CMs exist that are distinguished based on the output they produce: discrete and continuous. Discrete CMs output a class label (L) that indicates whether a word is correctly recognized (in the following L_{cor} is used to indicate the label for a correct hypothesis; L_{inc} to indicate an incorrect one). In the continuous case, the CM is a value ranging between 0 and 1 and is a measure of the quality of the underlying token (i.e., how confident is a speech recognizer that a word is hypothesized correctly). Table 4.1 shows an example of both CMs based on a hypothetically uttered sequence of words W and a possible recognizer output $\hat{W}$. For example, the word "quick" was incorrectly recognized and because of that should be assigned the label L_{inc} or a low CM value in the continuous case. The CM can then, for example, be used to detect this error and the ASR system re-evaluates the output hypotheses, or the user is asked for clarification.

Real utt.:	The	quick	brown	fox	jumps	over	the	lazy	dog
Class CM:	L_{cor}	L_{inc}	L_{cor}	L_{cor}	L_{cor}	L_{inc}	L_{cor}	L_{inc}	L_{cor}
Continuous CM:	*0.87*	**0.11**	*0.92*	*0.99*	*0.87*	**0.04**	*0.63*	**0.46**	*0.85*
Recognized utt.:	The	slow	brown	fox	jumps	under	the	crazy	dog

Table 4.1: A hypothetical example to illustrate how CMs identify correct and incorrect parts of a speech recognizer. It shows the hypothetical truly uttered sequence of words W, the hypothetical recognizer output $\hat{W}$, and the output of both types of CM, class-based and continuous. The mismatched words are underlined. Utterance is abbreviated "utt.".

Jiang (2005) and Seigel (2013) provide good overviews of previous studies in the area of CMs. Both classify the CMs into the following three categories:

1) *Utterance Verification (UV)*: this approach treats the problem of confidence estimation as a statistical hypothesis testing problem (using Likelihood Ratio Tests), where n which an approximation of the alternate hypothesis is needed for a reliable decision (cf. Sec. 4.3.2).
2) *Posterior probability approach*: this approach tries to calculate the true posterior probabilities by approximating the probability function that is normally dropped during the recognition process in the Maximum a-posteriori (MAP) rule (cf. Sec. 4.3.3).
3) *Classification approach*: here a statistical model is trained to estimate the CM in a post-processing step (cf. Sec. 4.3.4).

Each of the three approaches will be explained in more detail in the following sections.

4.3.2 Utterance Verification

Overview

In the UV approach, the CMs are estimated within the framework of statistical hypothesis testing, using the Likelihood Ratio Test (LRT). It answers whether the currently recognized word w, based on the acoustic observation O, has been recognized correctly and should be accepted as a valid hypothesis or recognized incorrectly and should be rejected. The two hypotheses are defined by (Seigel, 2013) as:

H_0: "w was correctly recognized and is generated by model λ_W"
H_1: "w was misrecognized and is generated by model λ_A"

where λ_W and λ_A denote the model for the null (H_0) and alternate (H_1) hypothesis respectively. Based on the definition of the null and alternative hypothesis, the LRT is defined as:

$$LRT(O, \lambda_W, \lambda_A) = \frac{P(O|\lambda_W)}{P(O|\lambda_A)} \underset{H_1}{\overset{H_0}{\lessgtr}} \tau \tag{4.1}$$

where τ denotes a threshold parameter to adjust the decision whether the null hypothesis is accepted or rejected. For simple, known probability functions of H_0 and H_1, under the Neyman-Pearson lemma (Neyman et al., 1933), the LRT is the most powerful test to decide which hypothesis is correct.

The general challenge in LRT is to formulate an exact representation of the alternate hypothesis in the denominator. Hence, for UV in ASR, the challenge is to build a reliable model λ_A that produces the desired distribution. For the estimation of model λ_A, different possibilities have been proposed. Rose et al. (1995) trained an anti-keyword model for each keyword of a twenty keyword recognition task. Sukkar et al. (1997) propose the combination of a non-keyword speech model and an anti-keyword model (misrecognitions) for a 10-digit recognition task. And as a last example, Rahim et al. (1997) use a combination of non-keyword (speech, background noise, silence) and an anti-keyword model for a 10-digit recognition task.

Applicability of UV

As stated above, the big challenge in UV is modeling the probability function of the alternate hypothesis λ_A. The exact estimation of this model is impossible in open and large-vocabulary speech recognition. Therefore, UV is only applied to closed-vocabulary problems like spoken digit or keyword recognition.

In the case of the Munich AUtomatic Segmentation System (MAUS), the UV approach could be considered if the phoneme domain was chosen for the prediction. In this domain, the phonotactic model that is generated by MAUS would correspond to the null hypothesis, and a combination of anti-keyword and background-noise models could be applied to model the alternate hypothesis.

In the current study however, the assessment should be performed on a word-, and not a phoneme-level. To implement this, language and acoustic models for Large Vocabulary Continuous Speech Recognition (LVCSR) would be needed. This is a problem for two reasons: first, in the MAUS framework, no such models are available and, second, building a reliable alternate hypothesis based on these would not be possible anyway (as mentioned at the beginning of this section). Therefore, the UV approach is not applicable to the

present study and in general to MAUS.

4.3.3 Posterior Probability Approach

Overview

In general, the posterior probability of Equation 1.2 would serve as a good metric to be used for a CM. Unfortunately, by dropping the normalizing term $p(X)$, the resulting posterior probability is not comparable across utterances and can not directly be used as a CM (the result of Equation 1.3 is not the real posterior probability $P(W|X)$, but more precisely the joint distribution of $P(O, W)$, cf. Sec. 1.3.2). As mentioned earlier in Sec. 1.3.2, $p(X)$ is impossible to calculate exactly and even hard to approximate, due to the many alternatives that have to be taken into account. Formally, it would be computed by (Jiang, 2005):

$$p(X) = \sum_{H} p(X, H) = \sum_{H} p(H)p(X|H) \tag{4.2}$$

where H denotes any possible hypotheses for the acoustic observations X. As the summation has to be done over all possible hypotheses, this would mean that all combinations of speech sounds (words, phonemes, hesitations, etc.) and non-speech sounds (noises, coughs, etc.) would have to be explicitly modeled (in most traditional ASR systems in a 39-dimensional continuous feature space). Therefore, this is computationally unfeasible (Jiang, 2005; Seigel, 2013; Pfister et al., 2008, p. 328).

CMs in the posterior probabilities class try to circumvent this problem by approximating the true posterior probability based on an approximation of the distribution of $p(X)$. The two most popular techniques for estimating $p(X)$ are the filler-based and the lattice-based approach.

The filler-based approach tries to estimate the desired normalizing distribution $p(X)$ by employing general filler or background models. An example of this is the use of an all-phone recognition, in which the score of an all-phone-recognition model, only constrained by the bigram probabilities of the language model, is subtracted from the score of the word recognition (Young, 1994). A second example would be the use of a catch-all model, as

done by Kamppari et al. (2000), in which the term $p(X)$ is estimated by summing over all available diphone model probabilities of a reduced diphone model. To create the reduced diphone model, the Gaussians are merged in a bottom-up clustering in which the two most similar Gaussians are merged until the model is small enough to be efficiently computable during recognition. A further example of a filler-based method is normalization based on the highest Viterbi word score output by the recognizer (Cox et al., 1996).

The lattice-based approach, as the name suggests, uses the compact representation of a high number of alternative hypotheses in a combinatorial way, as it is done in lattices. One example uses the sum over the probabilities of all possible hypotheses in the lattice as a normalization factor. A confusion network[1], a timeless, even more compact representation of the alternate hypotheses of a word sequence $\hat{W}$ than lattices, is then used to find the final best hypothesis for $\hat{W}$ (Mangu et al., 2000). As Mangu et al. (2000), Evermann et al. (2000) use the sum over the probabilities of all possible hypotheses in the lattice as a normalization factor and, additionally, modify the MAP term to include the product of the posterior probabilities of the current path. This acts as a local consistency measure to reward the connections that are supported by high-scoring alternatives. In another study, carried out by Rueber (1997), the author uses the recognizer probabilities from an N-Best list directly, after applying a re-normalization to the N-Best list so the probabilities sum up to 1. As a final example, Wessel et al. (2001) use a word-graph (that closely resembles a lattice in this particular case) to calculate the final posterior probability for word $[w; \tau, t]$ (τ is the start time and t is the end time of the current word) based on the maximum posterior probability between time τ and t of all alternative, overlapping hypotheses of the same word at a certain time point.

[1] A confusion network is based on two clustering steps: an intra-word step, where all words W_i that are identical and are overlapping are merged, and an inter-word step, where all words that have a phonetic similarity are clustered. Due to the structure of a confusion network, all initial hypotheses are retained. However, by only looking at the confusion network and following its different paths through the network, it is possible to create sentence hypotheses that were not present in the lattice representation (Mangu et al., 2000; Jurafsky et al., 2009, p. 374).

Applicability of Posterior Probability Approach

The filler-based approach as proposed by Young (1994) would technically be possible within the MAUS framework. This is because MAUS does provide free phoneme recognition. However, MAUS's free phoneme recognition does not result in good recognition in most languages (an exception being, e.g., Italian). It is not clear whether the performance is sufficient to achieve good normalization. An additional problem would be the alignment of the phoneme sequences, detected by MINNI, to the phoneme sequences that are output by Grapheme-to-Phoneme (G2P) conversion (and are used in MAUS). For this non-trivial task an alignment strategy would have to be found, which is likely to produce errors and would add noise to the data (an example for a tool that could provide this alignment is called TextAlign[2]; Reichel, 2012).

The approaches of Kamppari et al. (2000) and Wessel et al. (2001) and other lattice-based approaches are based on a high number of alternative paths. They are based on some combinatorial representation of many different paths that are orders of magnitudes higher, than those available in MAUS. The number of alternative paths in the MAUS lattice of mostly result from different start and end times of phonemes. This means that measures based on posterior probability are not applicable when used in combination with MAUS in its current form, as the lattice representation is different.

4.3.4 Classification Approach

General Overview

Classification-based approaches determine the confidence measure CM by learning a mapping function g based on a feature vector F containing n features $\{f_1, f_2, ..., f_n\}$ to output a label $L(F)$. Depending on the approach, the features are mapped directly into the according classes (correct – L_{cor} vs. incorrect – L_{inc}) by g_{cla}:

$$L(F) = g_{cla}(F), L(F) \in \{L_{cor}, L_{inc}\} \tag{4.3}$$

[2]This tool is also available as a web interface and service via `http://hdl.handle.net/11858/00-1779-0000-0028-421B-4`

Another possibility is to predict the probabilities of the respective class, which means that a value between 0.0 and 1.0 is output. This value then has to be interpreted. This can sometimes be done as easily as, for example, using a threshold τ to map the continuous values to the same labels as in the classification. Formally, a function g_{reg} is learned that maps the features F to the respective class label by using the threshold τ:

$$L(F) = g_{reg}(F) \underset{L_{cor}}{\overset{L_{inc}}{\lessgtr}} \tau, L(F) \in \{L_{cor}, L_{inc}\} \tag{4.4}$$

Classification approaches differ mainly in:

- type of ML algorithm (C4.5 tree, Artificial Neural Network (ANN), SVM, etc.)
- features used (decoder-based features, prosodic features, etc.)

In the following, both are described in more detail, followed by a discussion of the most relevant studies for the chosen approach.

Related Work – Machine Learning Algorithms

Many different algorithms have been used to estimate confidence measures (categorical or continuous). Find below a non-exhaustive list of examples of these methods used:

- *Naïve Bayes classification* (Zavareh et al., 2013)
- *Decision trees* (Schaaf et al., 1997; Kemp et al., 1997; Zavareh et al., 2013)
- *Boosted decision trees* (Stoyanchev et al., 2012)
- *Linear Discriminant Analysis (LDA)* (Schaaf et al., 1997)
- *ANNs* with
  - 3 layers; one node in the hidden layer and shortcut connections (Schaaf et al., 1997; Kemp et al., 1997)
  - 3 layers; 6 or 8 nodes in the hidden layer depending on the language (Tam et al., 2014)
  - 3 layers; 50 nodes in the hidden layer (Zhang et al., 2001)
  - 4 layers; 50 nodes in each of the two hidden layers (Weintraub et al., 1997)

- *Conditional Random Fields* (Pellegrini et al., 2010; Parada et al., 2010; Chen et al., 2013; Seigel, 2013; Ghannay et al., 2015)
- *SVMs* with
  - *Radial Basis Function (RBF) kernel* (Zhang et al., 2001; Xue et al., 2006),
  - *linear kernel* (Zhang et al., 2001; Zhou et al., 2004; Zhou et al., 2006)
  - *dot, polynomial, sigmoid, and ANOVA kernel* (Zhang et al., 2001)
- *RFs* (Xue et al., 2006).

Related Work – Features

The predominant class from which features have been proposed in the literature are decoder-based features. These features are either a by-product of the speech recognition process, for example, Viterbi decoding (Viterbi, 1967), or are designed based on the information available to the decoder. Popular examples of such features are the different parts of the reduced MAP decision rule (cf. equation 1.3), such as the language model score for a certain word w_i (Weintraub et al., 1997; Gillick et al., 1997; Xue et al., 2006), the acoustic model score for word w_i (Schaaf et al., 1997; Pellegrini et al., 2010; Xue et al., 2006), or the posteriors (or more precisely the joint probability $P(O, W)$) of the recognizer itself (Kemp et al., 1997; Gillick et al., 1997; Pellegrini et al., 2010; Stoyanchev et al., 2012; Zavareh et al., 2013; Tam et al., 2014).

In the past, different ways to normalize the three parts of the MAP rule have been proposed. For the two input components (language model and acoustic model), of the MAP equation, examples of such modifications are the maximum language model score in the N-best list (Zhou et al., 2004; Zhou et al., 2006), the acoustic model score normalized by a phone-only decoding[3] (Zhang et al., 2001), or the range and the minimum of the acoustic score of a word in an N-best list (Zhou et al., 2004; Zhou et al., 2006).

Similarly, several normalizations and modifications of the posteriors have been proposed. Examples of these kinds of features are the posterior score of the word divided by the prior (language model) probability of the word (Schaaf et al., 1997), the posterior

[3]The normalization is performed in a similar way to Young (1994).

score of the word divided by an approximation of $p(X)$ (approximation to the real posterior probability like in Sec. 4.3.3; Zhou et al., 2006; Weintraub et al., 1997), calculating the word posterior scores based on lattices (Chen et al., 2013), calculating the word posterior score based on confusion networks (Xue et al., 2006), and adding the scores of the previous and next word to make a prediction (Tam et al., 2014).

The list of examples of proposed decoder-based features is long. Some of them are acoustic stability, which is the number of times a word occurs in the word sequence hypothesis with different settings of the language and acoustic model weights (Schaaf et al., 1997; Kemp et al., 1997), the number of times a language model-backoff to a lower n-gram occurred (Schaaf et al., 1997; Zhang et al., 2001), and the order of the n-gram used (Weintraub et al., 1997). Further ones are the number of active final words states during the time segment T_W of the word (Schaaf et al., 1997; Pellegrini et al., 2010; Gillick et al., 1997), the hypothesis density that reflects the number of alternative links at different points in the word (beginning, end, and average over the complete segment) in the lattice (Kemp et al., 1997), and the N-best homogeneity, which equals the ratio between the best score in the N-best list and the total of the path scores of the N-best list (Zhang et al., 2001).

Besides the big group of decoder-based features, there are two smaller groups: syntactic and prosodic features. Examples of syntactic features are Part of Speech (POS) tags for the current word (Zavareh et al., 2013; Ghannay et al., 2015; Chen et al., 2013; Stoyanchev et al., 2012) and a binary feature that indicates whether the current word is common or not[4] (Zavareh et al., 2013). Examples of prosodic features are word duration as used in Schaaf et al. (1997) and Gillick et al. (1997) and speaking rate as in Schaaf et al. (1997).

Another often used feature is some version of word length. Examples of these are the number of phonemes in a word (Weintraub et al., 1997; Pellegrini et al., 2010), the logarithm of the number of phonemes in a word (Schaaf et al., 1997), or the number of letters in a word (Ghannay et al., 2015). Using length related features is based on the notion that shorter words are more difficult to recognize, than longer words (Weintraub

[4]Based on a list of "stop words", which are filtered out, for example, by search engines as they are too common to presumably be relevant in a search query. Examples of those words are "an", "by", "is", "it", "one", and "we" (for a complete list cf. Zavareh et al., 2013).

et al., 1997; Young, 1994).

Apart from the already mentioned features, a number of other features have been proposed that have not found widespread acceptance but are still worth mentioning. A few examples are the signal-to-noise ratio (Schaaf et al., 1997), the parsing mode, which indicates whether a word conveys semantic information (Zhang et al., 2001), the slot backoff mode for the semantic parser of a dialog system (similar to a language-model backoff; Zhang et al., 2001), the number of alternative candidates in a confusion network slot (Tam et al., 2014; Chen et al., 2013), and the edit distance between the hypothesized outputs of both an ASR system and a sub-word ASR system (Chen et al., 2013). Two further interesting examples are a heuristic bigram hit feature that reflects the number of hits a bigram gets in a search engine (exact search for the specified bigram; Pellegrini et al., 2010) and a topic feature based on the notion that erroneously hypothesized words do not fit the overall topic of a hypothesized sentence (or across utterances; Pellegrini et al., 2010).

Applicability of the Classification Approach

Some of the decoder-based features are available in MAUS. For example, the language model and the acoustic model probabilities, whereas other features that are based on lattices and confusion networks can, unfortunately, not be used, as their generation would drastically increase the modeling and execution effort.

An advantage of classification-based measures is that they allow the granularity of the decision to be changed. This means, even though features are based on the phone-level, the decision can be made on the word-level, and if the features are based on the word-level, the decision can be made on the sentence-level, etc. (as, e.g., done in Zhou et al., 2004; Zhou et al., 2006).

When assessing transcription quality, it is worth noting that MAUS has an advantage over an ASR system. In an ASR system, the same features are used for the estimation of the original word hypothesis and the estimation of the quality of this hypothesis. This is somewhat circular, as if these features were to strongly indicate a wrong ASR hypothesis, this might already have led to the recognizer outputting something different. In MAUS's alignment scenario, the transcription that is to be aligned with the speech signal is gener-

ated by human transcribers (or by an ASR system that most likely uses a different type of modeling). This means that the orthographic input into MAUS can be seen as a semi-independent knowledge source. It is only semi-independent as the transcription produced by the human labeler or an ASR system, is still dependent on what is being said in the input speech signal and on how well the input signal can be processed, for example, this might be difficult in cases in which the signal quality is bad. It is interesting to see that, after all, in ASR systems an estimation of the quality of the word hypotheses may use exactly the same features that are used for the original decision and still predict the quality of this decision in a reliable manner.

4.3.5 Classification Approaches: Relevant Work

Given the broad overview in the last part of this section, a small selection of studies, which are directly relevant to the current study, are explained in more detail in the following section.

Schaaf et al. (1997) pursue an approach that uses features that are extracted from the decoder. Examples of these are *acoustic stability*, which measures how often words occur when changing the weighting balance between language model and acoustic model, *language model backoff*, which describes the number of times the language model has to switch to a lower n-gram model, and the *normalized word score*, which is the word score calculated by the Viterbi decoder divided by the prior word probability. Schaaf et al. (1997) use two methods to classify words into the groups correct and incorrect: LDA and an ANN (3-layers with one single unit in the hidden layer and shortcut connections). In a follow-up study, Kemp et al. (1997) examine the use of lattice-based features in a classification approach (where the features are computed by the forward-backward algorithm, for example, *hypothesis density* describing the number of competing branches in the lattice, and once again *acoustic stability*). In this study, an ANN and a C4.5 decision tree are trained (the ANN has the same topology as the one previously used in Schaaf et al., 1997).

Zhang et al. (2001) use two kinds of features: decoder-based and parser-based features. An example of a used decoder-based feature is the *normalized acoustic score*, which is the

ratio of the word score, obtained by a language-model based recognizer, divided by the score of a phone-only decoding similar to Young (1994). Further examples include the *language model backoff* described above, *N-Best-Homogeneity*, which is the ratio between the score containing the hypothesized word and the sum of the scores of all paths in the list. As the ASR system used in this study is integrated into a dialog system, the researchers had access to parser-based features, such as whether a word conveys semantic information, and the slot backoff-mode on a two-word window (similar to a language-model backoff). A decision tree was then applied (the splitting criterion is information gain or Word Error Rate), as well as a neural network with one hidden layer (50 nodes), and SVMs using different kernels (Dot, Polynomial, RBF, Sigmoid, and ANOVA). In regard to classification accuracy, SVMs using an RBF and an ANOVA kernel are reported to achieve the best performance.

Xue et al. (2006) also used decoder-based features (such as acoustic and language model score), as well as confusion network-based features, such as their proposed entropy measure. They employed Decision Trees (DTs), SVMs with RBF kernels and RFs to predict class labels. To assess the different features, they used the Variable Importance (VI) output of the RF, which, however, is not unproblematic (cf. Sec. 3.7.4). They reported that the RF outperforms the accuracy achieved by the decision tree, as well as that achieved by the SVM.

4.3.6 Confidence Measures in Corpus Analysis

Paulo et al. (2004) examined how ASR based confidence measures can be used to assess the quality of automatic segmentation of spontaneous speech. Alignment is carried out using a hybrid approach between a more robust classical Hidden Markov Model (HMM) aligner, which is, however, speaker-adapted, and a more accurate Dynamic Time Warping (DTW) aligner based on speech synthesis, which is speaker-independent. They employ a two-step process, in which a rough alignment is performed initially employing the HMM aligner, and then refined for the final alignment by the DTW aligner.

The features used for Paulo et al. (2004) are based on the distances of the internal representation of the alignment procedure. Two of these features, for example, are “the

variance of the mean distance between the features of the recorded signal frames and the synthesized speech signal over the alignment path for a given phone" (DTW) and the "mean distance between the features of the recorded signal frames and the phone model"(HMM; for more information on the features cf. Paulo et al., 2004).

To distinguish between good and bad alignment, a threshold is employed at an Overlap Ratio (OvR) of 75% (cf. Sec. 2.6.1 for more details on the OvR). An OvR that is bigger than 75% is classified as a good alignment and an OvR that is smaller or equal to 75% is classified as a bad alignment. By doing so the continuous regression problem is transformed into a two-way classification task.

Three different classification algorithms were examined: a regression tree, an ANN, and an HMM. For the HMM classifier, two distinct models were trained, one for aligned and one for misaligned speech. For this, the best achieved performance, which is averaged over all phoneme classes (including silence) was 0.7324 for precision and 0.7117 for recall.

Paulo et al. (2004) showed that confidence measures can be used to detect misaligned phones after forced-alignment. One drawback of the method is using a speaker-adapted HMM aligner. Due to this, it is unclear whether the applied features could be successfully used when not adapted to the speaker. Another obstacle hindering the use of this approach in the current study is that MAUS does not provide a DTW aligner, which accounts for 50% of the features.

4.4 Measure of Confidence for Corpus Analysis (MOCCA) – Chosen Approach

4.4.1 Overview

As previously mentioned, in order to prepare a speech corpus so it can effectively be used in research, some time-intensive, highly repetitive, and error-prone pre-processing steps (annotation, manual/automatic alignment) are necessary. Fortunately, the aforementioned studies suggest that these errors can be detected automatically. The present study wants to contribute to this body of work by trying to tackle two different problems:

1) find errors in the transcription process that were either produced by a human or an ASR (experiment 1)

2) find errors in the subsequent automatic S&L (experiment 2)

Two aspects are worth noting explicitly in this context. First, the granularity of the prediction. In the current study, the prediction of the quality of the transcription and the subsequent alignment is performed at the word-level (in contrast to, e.g., the phoneme-level). Second, the two predictions (transcription/quality of alignment) are, to some extent, related, as it is assumed that wrong/bad transcripts will lead to worse S&L. However, this relationship is ignored in the following study, and the two problems will be analyzed independently in two different experiments. The granularity for both experiments is at the word-level.

4.4.2 Features

From the aforementioned features, the ones were selected that can be extracted from the MAUS process without further modeling. This means, all features based on the decoder and features directly built on top of the input data of the alignment process. These features comprise a subset of features that have successfully been used in Schaaf et al. (1997):

logLM: the log prior language model probability, $logP(W)$

logAP: the log posterior probability as produced by the Viterbi decoder of HTK (Young et al., 2002), $log[p(X|W) \cdot P(W)]$

logAPNorm: the log posterior probability normalized by log prior probability: $logAPNorm = logAP - logLM$

Duration: the segmented word duration

SpkRate: the local speaking rate, calculated as the ratio of mean word length in the training data $MeanDur$ and $Duration$: $SpkRate = MeanDur/Duration$

logNPhones: the logarithm of the number of phonemes in the target word according to the S&L

The MAUS system models phones, not words (cf. Sec. 1.3.3). As words w_i have a variable number of phones n, it is apparent that for each word w, n different feature values **logLM**, **logAP**, and **logAPNorm** are produced. To circumvent this problem of variable length feature vectors, functionals of these features are used. This ensures a fixed length feature vector and, depending on the functionals, allows the resulting features to represent the dynamic of the feature values over the evaluated phone-sequence. The used functionals are (always calculated over the feature vectors of all phones):

$$\begin{aligned}
\text{Sum:} \quad & sum(x) = \sum_{i=1}^{n} x_i \\
\text{Mean:} \quad & \overline{x} = \tfrac{1}{n} \sum_{i=1}^{n} x_i \\
\text{Median:} \quad & med(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right) & n \text{ even} \end{cases} \\
\text{Range:} \quad & range(x) = \max(x) - \min(x) \\
\text{Variance:} \quad & Var(x) = \tfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
\text{Standard Deviation:} \quad & \sigma(x) = \sqrt{Var(x)} \\
\text{DCT coefficients 1-3:} \quad & C_k(x) = \sum_{i=1}^{n} x_i \cos\left[\tfrac{\pi}{n}(i + \tfrac{1}{2})k\right], \text{ for } k = 1, 2, 3
\end{aligned}$$

where n is the number of phones, x_i is the feature value of the $i-th$ phoneme of a given feature, and $max(x)$ and $min(x)$ are the maximl and minimal value of the features over all phones respectively for all the functionals above. This yields a feature vector of constant dimensionality $d = 30$ for each word.

4.4.3 Training and Test Data

Training and Test Strategy

MOCCA was trained and tested on recordings from two different corpora. For training and parameter tuning with a 10-fold CV, a subset of the Kiel Corpus (Kohler, 1996; John, 2012)

was used. In each fold, a certain speaker was either part of the test or part of the training set. All generated folds are approximately equal with regards to training set and test set size (stratification). For testing the best performing parametrization, the PhonDat2 (PD2) corpus was used. This corpus is a completely independent test corpus, as neither speakers, nor recording setting and environment was identical.

Performing the training of MOCCA on one corpus and the evaluation of the performance against an unseen speech corpus, recorded under different conditions, ensures that the results can be used across corpora. Therefore, the method will be applicable to other speech resources as well.

Data

Kiel Corpus (training set): The used subset of the Kiel corpus consists of 30 speakers who produce a total of 2225 utterances of semi-spontaneous speech in the appointment scheduling domain as well as map task recordings (Kohler, 1995; John, 2012).

In the scheduling domain, two subjects made a series of appointments. To ensure that the subjects engage in conversation, the task was made more difficult by providing a calendar already containing other appointments. An example of the appointment scheduling task is

"ich kann ab vierzehn Uhr"[5]

as an answer to whether an appointment at that time would be possible. This allows for a controlled environment while the elicitation mode is semi-spontaneous.

In the map task, two subjects had to talk about two maps that are different in some way. This leads to discussions about the contents, in which again semi-spontaneous speech is produced (for more information, cf. Sec. 3.5.1 and Anderson et al., 1991). An example is:

"dann gehst du jetzt von der Eisenbahnlinie weg"[6]

[5]English (translation): "I am available after 2 PM"

[6]English (translation): "move away from the railroad".

In both recording modes, scheduling and map task, the subjects engage in semi-spontaneous speech, while solving the task together. However, the content of the utterances is task-specific.

PhonDat2 (PD2) (test set): The subset taken from the PD2 corpus (The ASR Consortium, 1995) consists of read speech of 16 speakers producing 64 utterances each, which adds up to a total of 1024 utterances. The domain of the recording is a train information query task, in which subjects were asked to read texts that represent dialogs that could appear in a train information system. An example of a prompt the subjects were asked to read is:

"geht heute noch ein Zug nach Hannover"[7]

Notes on the corpora used: First, both corpora have a manually corrected orthographic tier, and a manual S&L which was produced by trained phoneticians and both contain German utterances produced by native German speakers. Further, the manual S&L was corrected by a second annotator.

Second, as mentioned above, both corpora contain task-specific speech of the participating subjects. Performing training and hyperparameter optimization, and final testing in different recording modes will show whether a model is heavily optimized towards a task (as it will not perform well on the independent test set if this is the case).

4.4.4 Machine Learning Algorithms

Overview: For both experiments, two different algorithms were tested. For experiment 1 SVMs and RFs were used and Support Vector Regression (SVR) and RFs were used for experiment 2.

SVM: SVMs were already introduced and used in the last experiment (cf. Sec. 3.7.1). As stated before, SVMs are able to fit non-linearly separating hyperplanes in the low

[7]English (translation): "is there another train to Hannover today"

dimensional feature space using the *kernel trick*. The generated maximally separating hyperplane can be controlled by the penalty set for misclassified instances (soft-margin classifier). To do so, the cost parameter C is used to control this penalty. The lower the cost, i.e., when misclassifications are penalized less, the smoother the resulting hyperplane.

The two best SVM kernels reported in Zhang et al. (2001) are a Gaussian RBF[8] kernel of the form:

$$k(u, v) = exp(-\gamma \|u - v\|^2) \tag{4.5}$$

and the ANOVA RBF kernel of the form:

$$k(u, v) = \sum_{k=1}^{n} exp(-\sigma(u^k - v^k)^2)^d \tag{4.6}$$

As the training of the ANOVA RBF kernel is quite time-expensive and Karatzoglou et al. (2006) report that ANOVA RBF kernels generally perform well in regression problems, it is only applied for the regression task in experiment 2. Hence, the Gaussian RBF kernel was applied in experiment 1 and 2; the ANOVA RBF kernel only in experiment 2.

As the SVM is sensitive to its hyperparameters, they were tuned by performing a standard grid search: in case of the Gaussian RBF kernel the parameters for cost C (values tested: $C = 0.001$, 0.01, 0.1, 1, 10, 100) and γ (values tested: $\gamma = 0.001$, 0.01, 0.1, 1, 10, 100) were tuned. In case of the ANOVA RBF kernel the parameters C (values tested: $C = 0.1$, 1, 10), σ (values tested: $\sigma = 0.1$, 1, 10), and degree d (values tested: $d = 1$, 2, 3) were tuned.

For training the SVM with the Gaussian RBF, the package *e1071* (Meyer et al., 2015) of the R Programming Language (R) was used, which itself utilizes the *LibSVM* library (Chang et al., 2011), a parallelizable implementation of SVMs. For training the SVM with the ANOVA RBF kernel, the package kernlab (Karatzoglou et al., 2004) was used. Two different packages were used, as e1071 does not support ANOVA kernels, but can be parallelized and training is, therefore, much faster, whereas kernlab does support ANOVA and Gaussian RBF kernels, but processing is only single threaded.

[8]Gaussian Radial Basis Function will be abbreviated with RBF in the following.

Random Forest: RFs were also already introduced in Sec. 3.7.1. Despite the fact that they are reported to be insensitive to their hyperparameters, the two most important ones were tuned. These are the number of trees to grow ($ntree$ = 50, 100, 200, 500) and the number of random features to consider at each split in the tree ($mtry = 5 \approx \sqrt{d}$, 8, $10 = \frac{d}{3}$; the values for $mtry$ were rounded to the next whole integer). The *Random Forest Generator" (ranger)* package of *R* was used for training, as ranger was the best performing RF implementation available in R at the time of writing (Wright et al., 2015).

Class Probabilities in Classification: Both classification algorithms can output class probabilities p, instead of only a binary label. The class probability is a measure of how confident the classification algorithm is about its decision. The output values range from 0 (representing 100% confidence about class A) to 1 (representing 100% confidence about class B).

The class probability can be used to distinguish between those cases in which the classifier was certain about the decision, or edge-cases in which the classifier had problems making a distinction (around $p = 0.5$). The final decision about the label $label(x_i)$ is achieved by applying a threshold τ. Values x_i below the threshold are labeled as 'bad', above or equal to the threshold are labeled as 'good'. The default value is $\tau = 0.5$. The final decision is therefore defined as:

$$label(x_i) = \begin{cases} 'bad' & p_i < 0.5 \\ 'good' & p_i >= 0.5 \end{cases} \tag{4.7}$$

where p_i is the probability of the predicted class of feature vector f_i and ($p_{i,bad}+p_{i,good} = 1$) for all observations.

In the following, this information is used to screen out cases, in which the classifier could not distinguish clearly between the two classes. The possibility to select between instances in which the classifier could make clear distinctions, and where it could not, is a useful parameter to provide a user of the method with.

More details about how output class probabilities are estimated in SVMs can be found in Chang et al. (2011) and, for RFs, in Malley et al. (2012).

4.4.5 Receiver Operating Characteristic

The class probabilities output by the classifier can be used to a) skew the distribution to one side or the other (i.e., by lowering the threshold more instances are classified as good or vice versa) and b) to drop the instances which could not be clearly put in to one class or the other (e.g., discarding instances with $0.4 < p_i < 0.6$).

A receiver operating characteristic (ROC) curve is used (cf. Hastie et al., 2013, p. 316) to visualize how classifier accuracy is influenced, if instances close to $p_i = 0.5$ are left out of the decision in experiment 1 (cf. 4.5.2). The axes of the plot are the true positive rate (TPR) and the false positive rate (FPR). The TPR is also called sensitivity or recall and was introduced in Equation 3.5. The FPR is defined as

$$\text{FPR} = \frac{f_p}{f_p + t_n} \tag{4.8}$$

where f_p denotes the number of false positives and t_n is the number of true negatives.

4.4.6 Resampling of Feature Vectors

Various methods, to improve classification by generating more training samples in unbalanced datasets exist (for an overview of different algorithms, e.g., cf. More, 2016). One popular method is the Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al. (2002) and involves the following steps:

1. Select a random example of the minority class[9] and select its k nearest neighbors from the dataset (e.g., $k = 5$, as in Chawla et al., 2002).
2. Select one of the k nearest neighbors at random and randomly select a position in the feature space between these two points.
3. Add this point together with the minority class label to the dataset.

Chawla et al. (2002) not only show that this oversampling strategy improves results during classification, but also that results improve further in case the oversampling of the minority class is combined with an undersampling of the majority class.

[9] In the following the class with fewer examples is called minority class and the class with more examples is called majority class.

In the current study, a resampling for regression, not a classification task is necessary. Torgo et al. (2013) have extended the SMOTE algorithm to handle regression tasks and call it the Synthetic Minority Over-sampling Technique for Regression (SMOTER). Assigning a synthetic output in the regression case is not as trivial as in the original classification-based SMOTE. This is because, by definition, in the classification case the newly created samples always lie within the area that has already been occupied by the minority class (as a point is chosen in between two existing samples). Hence, assigning the class label is trivial. However, in the regression case, the continuous target variable has to be approximated.

SMOTER is similar to SMOTE with regards to selecting the two feature vectors that should be used to generate a synthetic sample (i.e., randomly selecting one of the k nearest neighbors of a randomly selected sample). To assign a target value to this new sample, Torgo et al. (2013) propose using the weighted mean of the two samples and propose the inverse distance to the training samples acts as the weight. They show that using this strategy improves results for their example dataset.

This algorithm will be used in the current study to generate more samples in regions that are occupied by too few training samples to allow a prediction of equal error across the range of values.

4.5 Experiments and Results

4.5.1 Overview of the Experiments

The general setup of the experiments was as follows: test data consisting of the speech signal and the corresponding transcript were automatically processed by the S&L system MAUS. During this process, the features explained in Sec. 4.4.2 are extracted from the decoder and the input data. Based on these feature vectors, MOCCA estimated two measures assessing the quality of the process: a) it tagged each word of the input transcript whether it matched the speech signal or not (experiment 1) and b) it estimated the degree of overlap of each word via the OvR between the calculated segmentation and the ground truth (manual) segmentation (experiment 2).

Despite common practice, the input features values were not standardized (i.e., to have a mean 0 and a variance 1) before being fed into the SVM with an RBF kernel. Standardizing the features was tested, but led to slightly worse results than for the unscaled data.

4.5.2 Experiment 1: Correctness of Transcription

In experiment 1, estimating whether a word label in the transcription is correct is treated as a two-class classification task. Two different classifiers were tested, an SVM with an RBF kernel and an RF.

According to the related work in Sec. 4.3.4 and 4.3.6, decoder information is a sufficient knowledge source to determine the quality of an ASR and an automatic S&L process on the phone-level. Therefore, the current experiment should answer the research question: *Do the features described in section 4.4.2 carry enough information to classify each word in the input transcription into the classes 'correct transcript' versus 'incorrect transcript' at a level above that of chance after performing an automatic S&L using MAUS?*

As mentioned in Sec. 4.4.3, the corpora are hand-labeled and manually corrected by two people. It is assumed that no transcription errors exist in the corpus or if they do, they exist to such a small degree as to be negligible. Since examples of the incorrect class are needed for training, the following replacement strategy was applied to every test recording to artificially introduce this kind of error.

Transcription Errors: First, an automatic S&L using MAUS was performed on the test recording, and the features were extracted from the decoder output for all words. For all words that have an OvR of more than 90% between the MAUS S&L and the ground truth segmentation, the MAUS S&L was repeated, however the selected word was replaced by another, wrong word in the transcript and features were extracted from the decoder output for the replaced word.

The replacement word w_r was a randomly selected word from the corpus. To be considered as a valid replacement, it had to fulfill the following two constraints:

- the length, of the canonic representation of the word, of the replacement word w_r needed to be $length(w_r) = length(w_o) \pm 1$
- the word-length normalized Levenshtein distance (Levenshtein, 1966) had to be at least 75%

If no word could be found in the range ± 1, it was incrementally increased until a replacement was found. The length constraint was necessary, as it would have otherwise not always been possible to execute the alignment procedure. Due to a minimal length of 30 ms for each phoneme, constellations exist in which an arbitrarily selected phone-sequence cannot be put into a specific time segment. This leads to an error during the MAUS alignment. To circumvent this problem and allow a transparent replacement strategy, the length constraint was introduced.

An example of a rejected replacement would be "train" and "rain", as it only fulfills the length requirement, but not the Levenshtein requirement. A valid substitution would be the replacement of "train" by "wash". The "Levenshtein distance based replacement" constraint was introduced, as it was assumed that a wrong word which is (phonetically) similar, like "train" to "rain", is easier to align to the speech signal. This similarity, in turn, means that the feature values will not differ significantly and are, therefore, harder to detect. Additionally, an error like this would probably lead to fewer problems in the subsequent processing than a word that has no similarity at all (subsequent misalignments).

Employing a strategy like this has two benefits: first, many training examples can be generated automatically and, second, the training set is balanced with regard to the output classes (each relevant part in the speech signal is analyzed twice – once with a correct and once with a wrong transcript). This leads to a total of 26,649 training examples.

Results: Table 4.2 summarizes the results of the two-way classification, based on a hyperparameter search with the aforementioned 10-fold CV. The metrics accuracy, precision, and recall are reported for the models' hyperparametrization, yielding the best accuracy (cf. Sec. 3.7.2). The exact hyperparameter values are given in the caption of Table 4.2. The upper-half of Table 4.2 shows the results of the 10-fold CV; the lower half shows the

results when testing the best parametrization against the independent test set PD2.

Corpus	Class.	Accuracy	Precision	Recall
Kiel	SVM	0.7822	0.7897	0.7672
	RF	0.7908	0.7862	0.7968
PD2	SVM	0.7876	0.7785	0.7868
	RF	0.7526	0.6923	0.7794

Table 4.2: Performance of the SVM and the RF classifiers and according metrics. The SVM was built with hyperparameters $C = 100$ and $\gamma = 0.1$. The RF was built with $ntree = 500$, $mtry = 8$. For both SVM and RF a decision threshold of $\tau = 0.5$ was used.

It can be seen that the SVM and the RF classifier both achieve a similar performance in the 10-fold CV. The RF has a slightly better accuracy than the SVM, which agrees with the findings put forward by Fernández-Delgado et al. (2014).

When testing against the independent test set, the accuracy obtained for the SVM is close to the level of accuracy for the original dataset. It seems to generalize better in this case than the RF does. The RF, furthermore, shows a skewed distribution towards predicting more f_n (in the current case labeling more 'good' instances as 'bad'), which leads to a decrease in precision by more than 10%.

The ROC curves of the two best parametrizations of the SVM and the RF can be seen in Fig. 4.1. When looking at Fig. 4.1, it can be observed that the RF outperforms the SVM in the CV training across the range. When applied to the test set, the SVM outperforms the RF at all thresholds and it even performs slightly better on the test set than during the CV evaluation.

As mentioned in paragraph 4.4.4, both classification algorithms are able to output class probabilities. When observations, in which the probabilities indicate that the decision is not clear (values around 0.5, meaning that both classes are equally likely), are left out, the performance of the classifier improves. This can be seen in the respective ROC curve in Fig. 4.2. This could, for example, be used to only output instances, in which the classifier could make a clear decision (i.e., to filter out instances that are unclear, in case enough

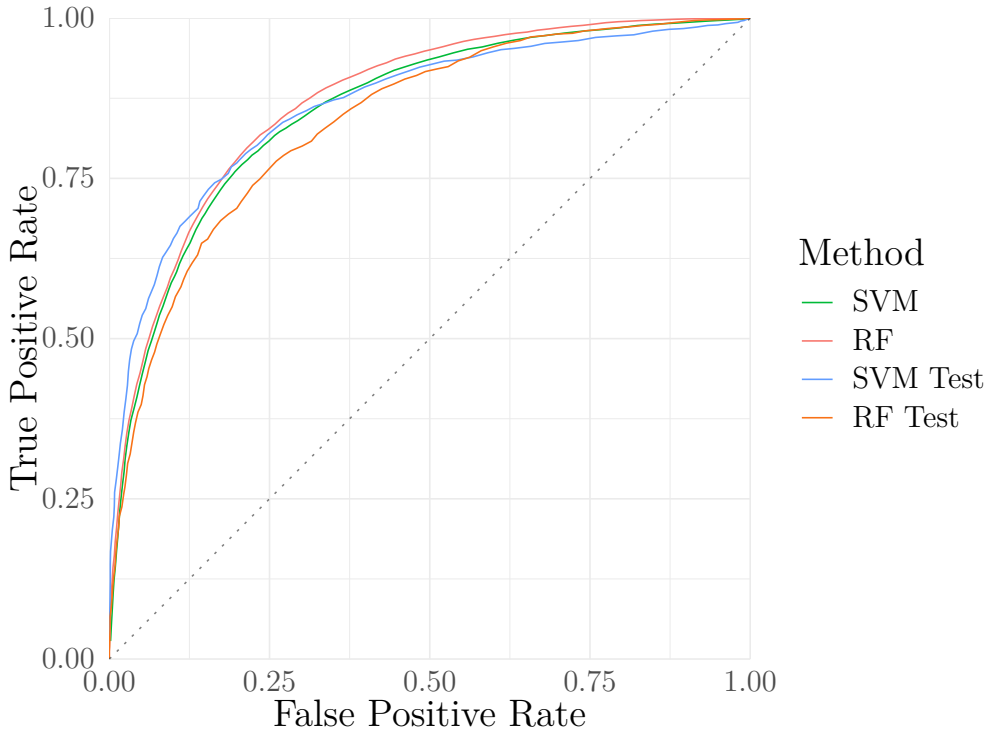


Figure 4.1: The four ROC curves of the best parametrization for the SVM and the RF when being applied to the training and the test set with varying threshold τ from Equation 4.4 (SVM CV: green; SVM test: blue; RF CV: red; RF test: orange).

input data is available). Five different gaps are examined which are 0.0 (original setting which is also shown in Fig. 4.1), 0.2, 0.4, 0.6, and 0.8. As an example, having a gap of 0.2 means that values that are predicted with a probability between 0.4 and 0.6, are left out of the evaluation (no output in those cases).

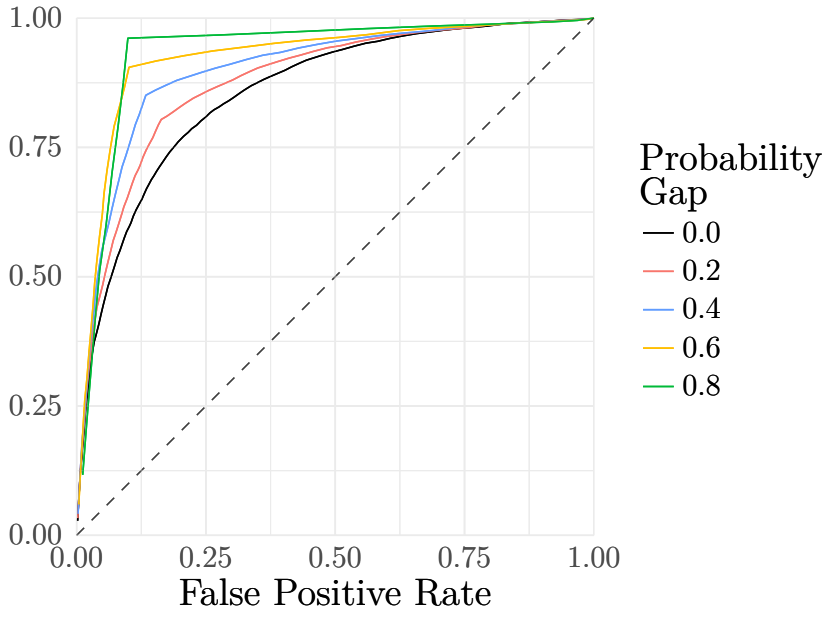


Figure 4.2: Four ROC curves showing the resulting performance of the SVM for a) varying thresholds τ from Equation 4.4 and b) leaving out instances predicted with certain class probabilities around the instances predicted with $p_i = 0.5$. For a gap of 0.2 this means that the label 'bad' is output for probabilities $p_i = 0.0...0.4$ and the label 'good' for probabilities between $p_i = 0.6...1.0$. Five different gaps are evaluated 0.0, 0.2, 0.4, 0.6, and 0.8.

Summary - Experiment 1: The results of experiment 1 are promising, when it comes to detecting transcription errors in speech signals. One problem though is the skewed distribution of errors and how they occur in speech corpus preparation. Unlike in the experiment, correct and incorrect words do not occur with a ratio of 1:1. It is assumed

that the incorrect:correct ratio is in the range between 50:1 and 500:1. In the experiment roughly 22% of words were labeled incorrectly. In an assumed case of a correct:incorrect ratio of 250:1, this means to detect 8 out of 10 errors in 2500 words, around 550 correct instances (that were falsely labeled "correct") have to be checked as well. Having to check around 550 words is, of course, better than checking all 2500 words, but still represents a high number of tokens for manual validation.

If a corpus is big enough, leaving out the cases in which the classifier is not certain about its decision, would increase the performance of the classification algorithm. This, in turn, would lead to less manual work when trying to find the remaining errors, but also means that some of the input data has to be discarded in further processing.

4.5.3 Experiment 2: Segmentation Quality

Overview

The experiments described in the following section investigate the prediction of the OvR between manually and automatically created S&L (cf. Fig. 4.3). As the OvR is a continuous value its prediction is a regression task. For those experiments, the algorithms used were again RFs (as in the last experiment) and SVR (cf. Sec. 3.7.1) with an RBF kernel. Zhang et al. (2001) report that the ANOVA Radial Basis Function (ARBF) kernel achieves good performance as well. This could not be confirmed in a 10-fold CV, as the predictions are only weakly negatively correlated to the actual values. Because of this and the long training times of the ARBF kernel, this was not further investigated and more detailed results will not be reported here.

As the literature suggests, it is possible to use decoder-based features to do an estimation of the quality of the automatic S&L process (Paulo et al., 2004). Even though it has previously only been done on the phoneme level, there is no evidence that this should be different for a quality estimation on a word-level. The ground truth for the OvR, which is used for training, is calculated between the segment boundaries of an automatic S&L produced by the MAUS system and a manually obtained S&L created by human transcription on the aforementioned corpora (Kiel corpus and PD2 corpus). Estimating this OvR

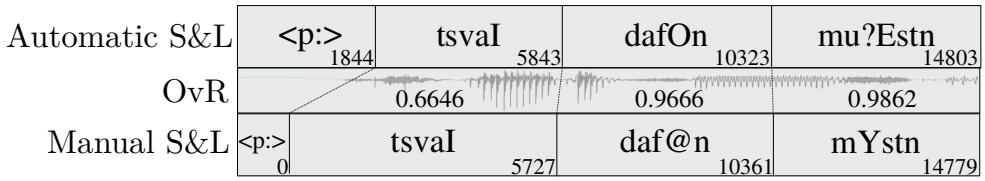


Figure 4.3: A real example of phoneme strings and their alignment: an automatic S&L (top), a manual S&L (bottom), and the resulting OvR values (middle). Additionally, the time index in samples is shown as extracted from the signal (zeroed at the first relevant sample), where the numbers belong to the boundary on the right side of it.

value during the segmentation process (where no human ground truth is available) would allow possible transcription errors to be pointed out right after processing and would save time in correcting erroneous S&L. This leads to the following research question for the current experiment: *Do the features described in section 4.4.2 carry enough information to approximate the OvR between an automatic S&L and a manual S&L?*

It is worth noting that an alignment between two segments, in which the end time of the one segment is equal to the start time of the other segment, is as bad as it would be if there was no gap between the two segments. As the OvR can have smaller values than 0 (cf. Sec. 2.6.1), all values that are smaller than 0 are set to 0.

Analogous to experiment 1, only the results based on the parametrization that achieved the best performance with regard to the correlation coefficient (Pearson) are reported. Reported are therefore: the hyperparameters tuned and the two measurements *correlation coefficient* (Pearson) and *mean absolute error (MAE)* for the best parametrization (for the metric definitions, cf. 3.7.2).

The original distribution of the OvR can be seen in Fig. 4.4 (gray). It can be seen that the ground truth OvR values are not equally distributed. The dense regions around an OvR of 1 might lead to an overfitting of the regression algorithm in that area and, therefore, resampling strategies were investigated. This lead to three different experiments:

a) Training the regression algorithm using the full dataset, including dense regions (experiment 2a),
b) Training using a dataset resulting from an undersampling of the bins with many

observations (experiment 2b)

c) Training using a dataset resulting from both oversampling of the bins with only a few observations and undersampling of the bins with lots of observations (experiment 2c)

The histograms of the distribution of the OvR values in the three experiments 2a, 2b, and 2c can be seen in Fig. 4.4.

An undersampling strategy was used in two of the experiments. In these experiments, a subset of the original dataset was used to build a model that predicted the OvR with a more equal accuracy across the range of possible values. For this, the OvR range of $[0, 1]$ was divided into 20 equally sized bins all of which had a width of 0.05. Bins that contain more observations than 1890 observations (the mean number of observations calculated over all bins) are called "majority bins" (e.g., the interval $0.9 - 0.95$ in Fig. 4.4), bins that contain fewer observations than 1890 are called "minority bins" (e.g., the interval $0.10 - 0.15$ in Fig. 4.4) in the following. Those names agree with the definition of minority and majority classes in class-based SMOTE (Chawla et al., 2002).

A 4th experiment deals with the different ways in which two time segments can overlap. The different types of overlays possibly make the estimation more difficult, as it changes the dynamic shape that the features have to capture. To test whether this makes a difference to the prediction performance, an experiment is conducted in which this information is fed into the SVR. This is an experiment, to see whether this information can be used to improve the prediction. However, it can not easily be used in a real application, as this feature is not available and would have to be estimated as well. The results for this experiment can be found in App. B.1.

Estimating the OvR – Experiment 2a

In this experiment, the unaltered, complete dataset was used during the training phase. This means that for the 10-fold CV a total of 55,515 observations from the Kiel corpus were used.

The results are summarized in Table 4.3. It can be seen that the RF slightly outperforms the SVR. However, both algorithms are comparable performance-wise. As mentioned before, Fernández-Delgado et al. (2014) report that RFs and SVMs often perform similarly

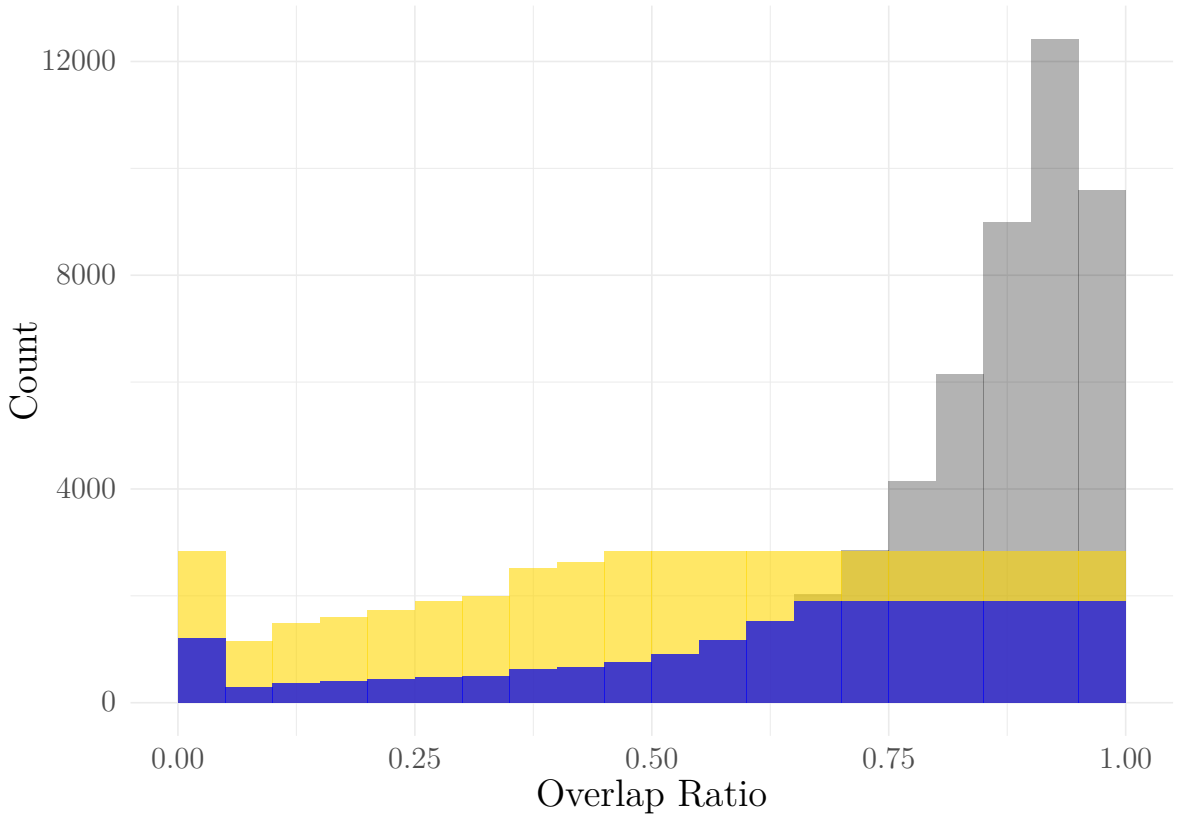


Figure 4.4: Histogram showing the original distribution of the overlap ratio (light gray), the undersampled dataset (blue), and the dataset that was oversampled in the minority classes and undersampled in the majority classes (yellow). In the undersampled dataset, each bin contains the average number of observations calculated over all bins of the original dataset (1890). In the over-/undersampled dataset each majority class contains 1.5 times the average number of observations calculated over all bins (2835) and minority classes are oversampled by 300% (but are not allowed to contain more than 1.5 times the average of observations).

in classification tasks, which seems to also hold true for regression tasks in the current case.

When applied to the independent test set, the RF predicts values with a higher correlation between real and predicted values when compared to the SVR. The best parametrizations based on the grid search were for the SVR $C = 1$ and $\gamma = 0.1$, and for the RF $mtry = \frac{1}{d}$ and $ntree = 500$.

Moreover, it can be seen that the correlation coefficient decreases and so does the MAE. This seems counter-intuitive, but there are cases in which the correlation coefficient decreases when the prediction occurs more frequently in a different direction than the real values, the deviation to the real values, however, is smaller than before.

Corpus	Class.	CorCoeff	MAE
Kiel	SVR RBF	0.7337	0.0918
	RF	0.7487	0.0949
PD2	SVR RBF	0.6558	0.0783
	RF	0.6622	0.0809

Table 4.3: Results of the best parametrizations according to the Pearson correlation coefficient (CorCoeff) of the two classifiers (Class.) SVR with Gaussian RBF (RBF) kernel and the RF. The RBF SVR was built with parameters $C = 1$ and $\gamma = 0.1$; the RF was built with parameters $ntree = 500$ and $mtry = \frac{d}{3}$.

Fig. 4.5 shows the distribution of the predicted values versus the original values for the best SVR parametrization. For this visualization, the results of the SVR are used, as it outperforms the RF in the following experiments and making the plots more comparable across experiments. The plots for the best RF look similar. It can be seen that the prediction is more accurate close to OvR 1, compared to OvR 0. This is because the OvR is not equally distributed over the possible range of values.

As can be seen in Fig. 4.4, there are many values close to 1 (light gray bars), indicating an almost total overlap between automatically estimated word segments and the ground truth segmentation, and few values < 0.5, indicating a small overlap. While this is the desired behavior, as this means the MAUS S&L is mostly correct, it makes it difficult to

train a model that works equally well over the complete range of values. The half-violin plots[10] in Fig. 4.5 suggest that the error for the OvR prediction is heteroscedastic (bigger close to 0 and smaller close to 1).

In the context of MOCCA especially, a prediction that has an equal error distribution over the complete range of predicted OvR values is of high importance. This becomes even more important when badly aligned segments are of special interest. Currently, the badly aligned segments cannot be predicted reliably.

Undersampling of Majority Bins – Experiment 2b

In this experiment, the majority classes (cf. Sec. 4.5.3) are undersampled to prevent the regression algorithm from favoring dense regions with many measurements during the training phase (as happened in experiment 2a). The undersampling strategy limits the number of observations in each majority bin to an arbitrary number of observations, which is the mean amount of observations calculated over all bins in the current case. This limit is 1890 observations. Therefore, if more than 1890 observations fall into a bin, the correct amount of observations is randomly selected from the available measurements. This equals an undersampling strategy, in which the classes with a higher number of observations are undersampled to a greater extent than the bins with fewer observations. This strategy results in a total of 22,552 training observations in the Kiel corpus (of the original 55,515 samples; this equals a loss of 60% of the data).

Table 4.4 summarizes the results of the hyperparameter search for the SVR and the RF. It once again can be seen that the results look similar when comparing the SVR and RF regression. Analogous to experiment 2a, in which the model is applied to the independent test set, the correlation decreases for both classifiers, an indicator that the models do not generalize well. In the current study, the SVR generalizes better than the RF. Overall, the results are slightly worse than in experiment 2a. This was to be expected, as the

[10]Half-violin plots are used, as boxplots do not show the distribution of values over the range. Using half-violin plots should save space and avoid graphical problems in visually parsing the distribution by the symmetry in violin plots as suggested by Irizarry, 2017. By marking the positions of the 25%, 50%, and 75% quartiles it combines the advantages of both violin plots and boxplot.

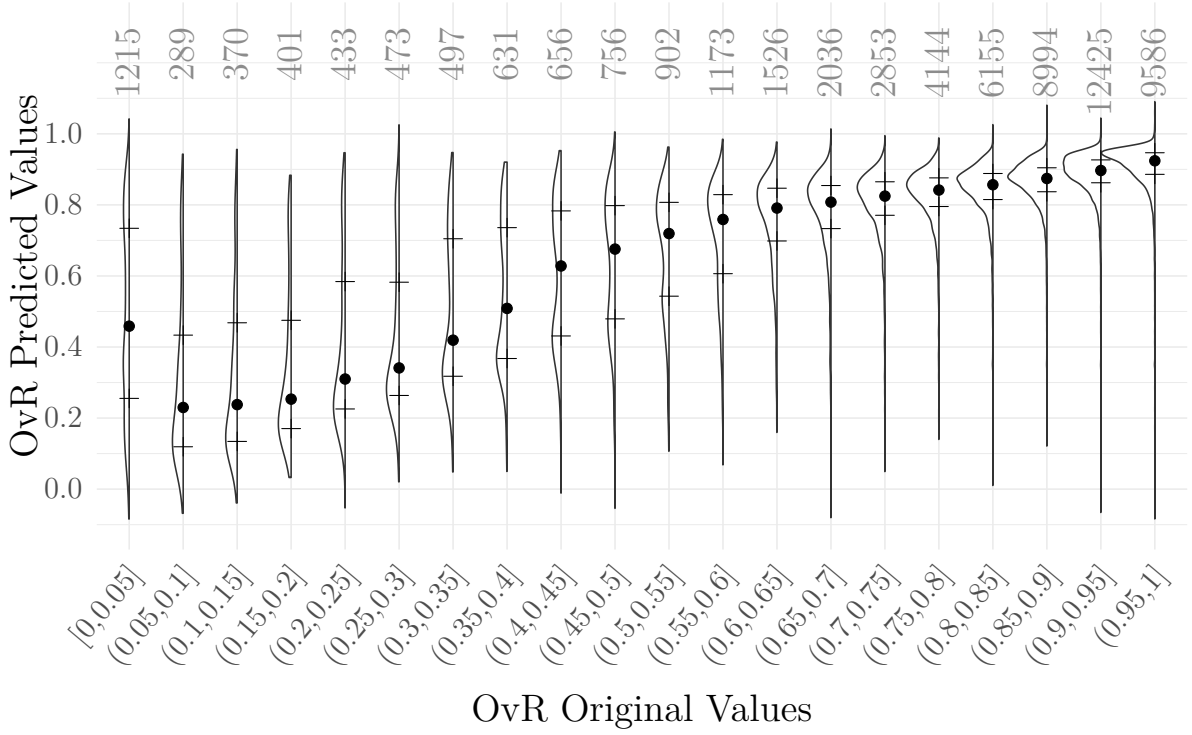


Figure 4.5: Visualization of the original and the predicted OvR values for the best SVR parametrization for the original/unbinned dataset. The half-violin plots show the variation within each bin. The number of observations in each bin is plotted above each plot. '[' and ']' on the x-axis indicate that the boundary value is part of the interval, '(' and ')' indicate that the value is not part of the interval. The black horizontal lines in the half-violin plots (-) indicate the 25% and 75% quartile; the • the median.

test set also contains more OvR values close to 1. By training the model in a way that predicts the OvR across the range of values better, many values close to 1, that benefitted from the skewed prediction before, lead to a decrease in overall performance. Considering the amount of data that was removed from the upper bins for training, the result is still promising.

Corpus	Class.	CorCoeff	MAE
Kiel	SVR RBF	0.7367	0.0918
	RF	0.7430	0.1290
PD2	SVR RBF	0.6415	0.0946
	RF	0.5955	0.1143

Table 4.4: Results of experiment 2b for the best parametrizations according to the Pearson correlation coefficient (CorCoeff) of the two classifiers (Class.) SVR with Gaussian RBF kernel and the RF. The RBF SVR was built with parameters $C = 10$ and $\gamma = 0.1$; the RF was built with parameters $ntree = 500$ and $mtry = \frac{d}{3}$.

The half-violin plots show (cf. Fig. 4.6) that the bins close to 0 lose some of their variability and bins close to 1 gain some. Additionally, the predicted values close to 0 have a lower median, which means that the prediction in these bins moved closer to the actual ground truth OvR values, with the first bin posing an exception. This bin not only contains the values between 0.00 and 0.05, but, due to setting all OvRs < 0 to 0, all originally negative OvRs as well. This might be the reason why the values in this bin were difficult to predict. Furthermore, it can be seen that the OvR values in the majority bins close to 1 lose some of their accuracy when predicting. Overall, it seems that the prediction performs better over the range of values with the undersampling strategy.

Oversampling/Undersampling Experiment 2c

In this experiment, the previous experiment 2b is extended by an oversampling of the minority bins (cf. Sec. 4.5.3). This should allow the model to take more information into account from majority bins close to an OvR of 1 (which were undersampled massively in

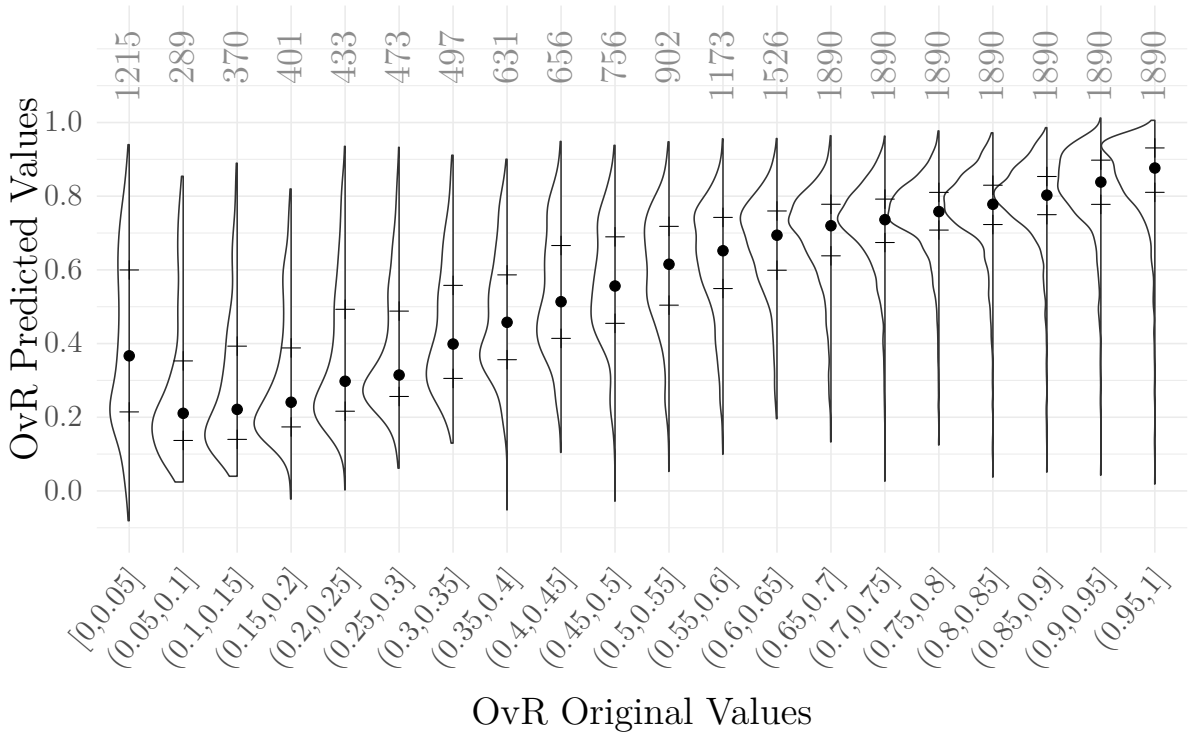


Figure 4.6: Visualization of the original and the predicted OvR values for the best SVR parametrization for the undersampled dataset. The original values are put into bins of the size 0.05 between 0 and 1. The half-violin plots show the variation within each bin. The number of observations in each bin is plotted above each plot. '[' and ']' on the x-axis indicate that the boundary value is part of the interval, '(' and ')' indicate that the value is not part of the interval. The black horizontal lines in the half-violin plots (-) indicate the 25% and 75% quartile; the • the median.

the previous experiment).

In order to do so, the minority bins were oversampled by 300%, which is a common value for SMOTER (cf. Torgo et al., 2013), to increase the number of examples in these bins. Additionally, analogous to before, an undersampling of the majority classes was performed. In the current experiment, the majority bins were allowed to contain 1.5 times (arbitrarily chosen with the restriction that it has to be higher than in experiment 2b) the mean of observations calculated over all bins. This is possible, as the oversampling of the minority bins allows balancing the dataset by adding more instances in the lower bins and still keeping a balanced dataset.

To prevent bins in the minority classes having more instances after oversampling than undersampled majority bins, they are limited to contain 1.5 times the mean of the observations over all bins as well. This is required for all bins that fall into the minority bin category and contain more than $\frac{1}{3} \cdot mean$ observations, for example, bins 0.50, 0.55, and 0.65 (cf. Fig. 4.6). This strategy results in a total of 49,020 training observations (of the original 55,515 samples from the Kiel corpus, 28,368 were kept, which equals a loss of around 49% of the data compared to the original experiment 2a, but a gain of 10% when compared to experiment 2b; 20,652 observations were generated for the minority bins by SMOTER).

Table 4.5 shows that the performance of the RF improves more when applied to the training set than the SVR when compared to experiment 2b. It seems the RF makes better use of the larger amount of training observations. On the other hand, it also seems that the RF overfits the data in the training phase, similarly to experiment 2b, and the performance regarding the correlation coefficient on the test set decreases more than is the case for the SVR.

Similar to experiment 2b) it can be seen in Fig. 4.7 that the variation in minority bins close to an OvR of 0 is once again reduced and that the values move closer towards their original ground-truth values. This goes along with an expected increase in variation in the majority bins close to 1. Moreover, as was the case in experiment 2b, it can be seen that the variation in the first bin is still high.

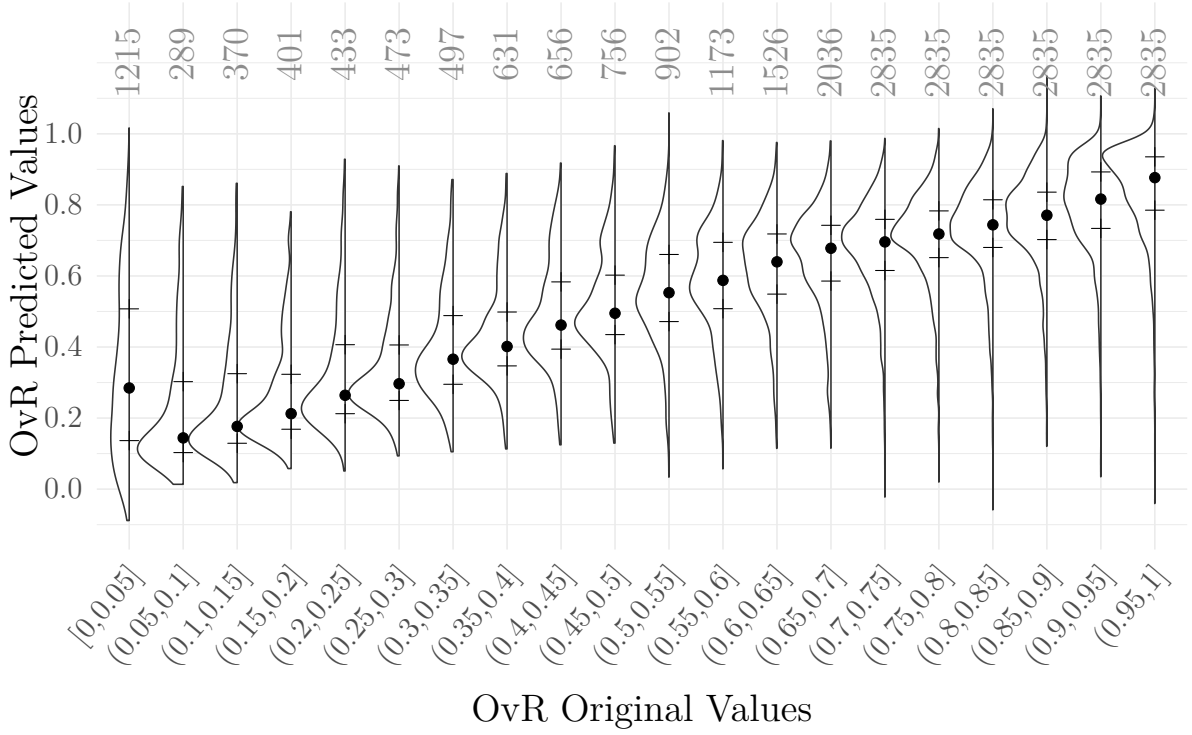


Figure 4.7: Visualization of the original OvR values and the predicted ones for the best SVR parametrization for the over- and undersampled dataset. The half-violin plots show the variation within each bin. The number of observations in each bin is plotted above each plot. '[' and ']' on the x-axis indicate that the boundary value is part of the interval, '(' and ')' indicate that the value is not part of the interval. The black horizontal lines in the half-violin plots (-) indicate the 25% and 75% quartile; the • the median.

Corpus	Class.	CorCoeff	MAE
Kiel	SVR RBF	0.7416	0.1259
	RF	0.7732	0.1183
PD2	SVR RBF	0.6006	0.1161
	RF	0.5992	0.1175

Table 4.5: Results of the best parametrizations according to the Pearson correlation coefficient (CorCoeff) of the two classifiers (Class.) SVR with Gaussian RBF kernel and the RF. The RBF SVR was built with parameters $C = 1$ and $\gamma = 0.1$; the RF was built with parameters $ntree = 500$ and $mtry = \frac{d}{3}$.

OvR prediction error against word length: As stated in Weintraub et al. (1997) and Young (1994), the CM is harder to estimate for shorter than for longer words. In Fig. 4.8 the MAE of individual predictions of the OvR in experiment 2c is plotted against word length (based on the number of phones in the canonic form of the word). It can be seen that this finding can be reproduced using the current dataset.

Summary – Experiment 2

Experiments 2a to 2c examined whether the OvR can be reliably predicted. It was found that the OvR can be predicted with a moderate correlation in the independent test set. By applying an undersampling and an undersampling/oversampling strategy the correlation coefficient slightly decreases. However, this leads to an increase in prediction accuracy, especially for bins close to 0.

The MAE of the OvR prediction, grouped by the bins of the three experiments 2a - 2c, is summarized in Fig. 4.9 (yellow: experiment 2a; blue: experiment 2b; green: experiment 2c). It can be seen that the error gets smaller in the lower bins for the undersampling and the undersampling/oversampling strategy, but slightly increases in the upper bins. The reason for this is the large number of instances in the bins close to 1. However, the error is more similar across the range of values. In the case of MOCCA especially, where errors in the automatic S&L should be detected, it is paramount to be able to detect values close to

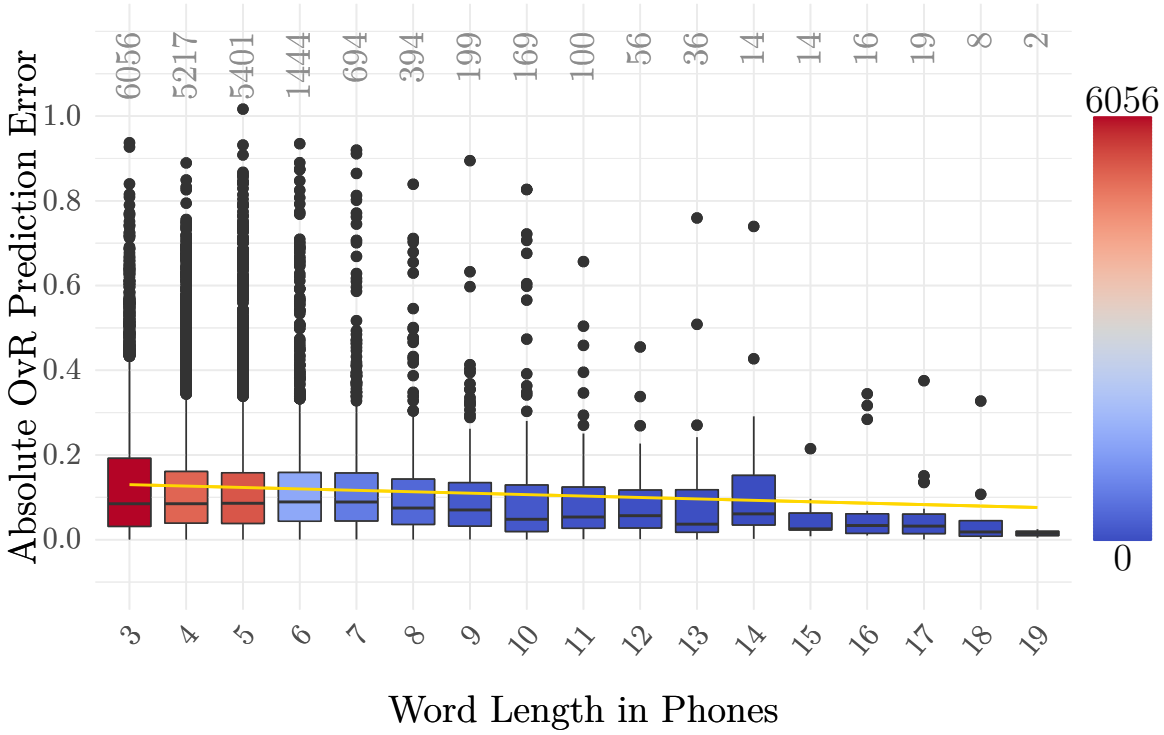


Figure 4.8: Overlap prediction error plotted against word length. The number on top of each boxplot indicates the number of observations in that bin, as does the color of the box. The yellow line is a linear regression line fitted to the data and indicates a downward trend, meaning that longer words are generally easier to predict than shorter ones.

0 in a reliable fashion. This means that the model with the undersampling/oversampling strategy is best suited for the current task.

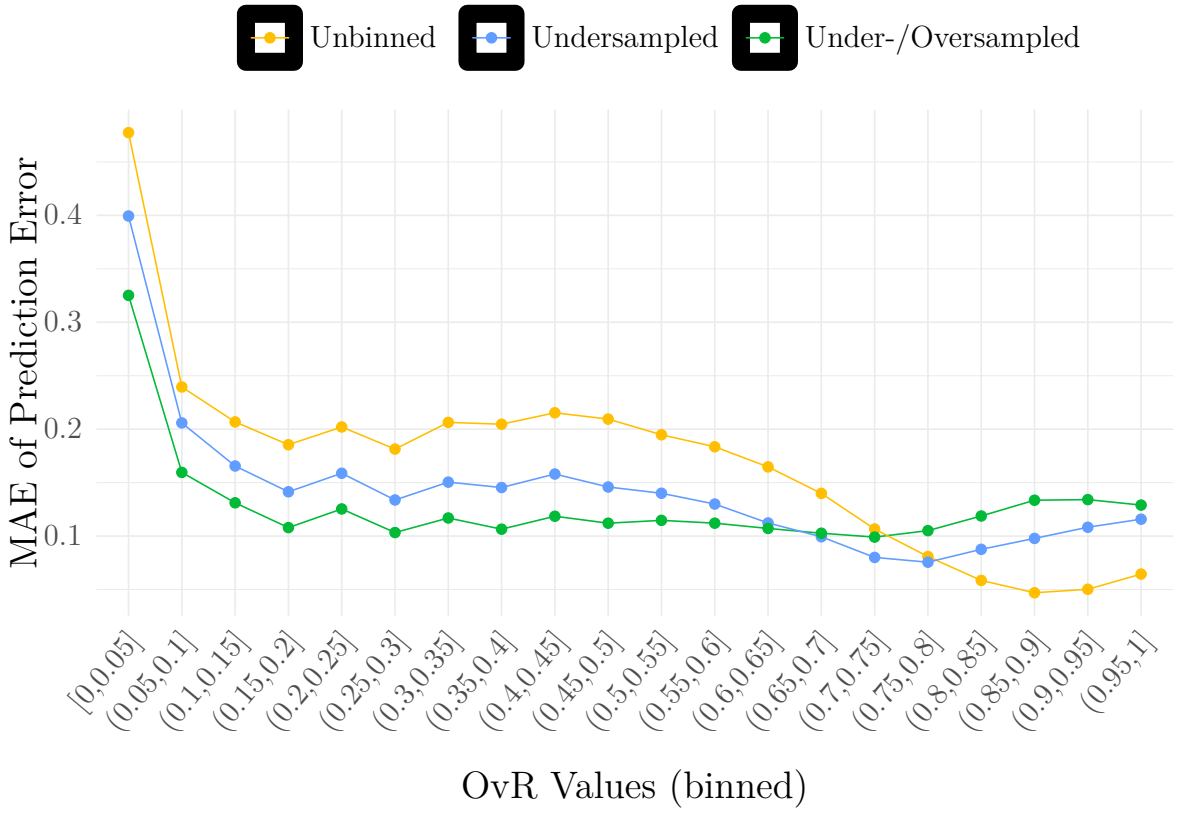


Figure 4.9: Each line shows the MAE of the prediction of the OvR in each bin between 0 and 1 (yellow: experiment 2a - unbinned/original data; blue: experiment 2b - undersampling strategy; green: experiment 2c - combined undersampling and oversampling strategy).

The advantage of the oversampling strategy is that the center of gravity of the histogram of OvR values can be shifted towards the center, even when using more data from the original dataset. This allows adding more information to the upper bins without skewing the predictive model towards the dense regions of the target variable.

4.6 Summary and Discussion

The outcome of experiment 1 shows that erroneous words in a transcript can be detected at a level above chance from the features extracted from the MAUS S&L alignment pro-

cess. The best model prediction accuracy is about 78% and produces roughly equal error types (i.e., false negatives and false positives are equally likely). The SVM is better at generalizing using the training data and outperforms the RF when applied to data from another corpus. This is an important aspect in real-world applications. Therefore, this should be taken into consideration if MOCCA is implemented as a web service as planned.

As mentioned earlier, the classification process implemented here has an advantage when compared to confidence measure estimation in ASR systems. The advantage is that two different, mostly independent knowledge sources are combined. Namely, information from the transcriber (be it a human or an independently modeled ASR system) and information extracted from the MAUS alignment process.

Experiment 2 demonstrated that the model best suited for a real-world application predicts OvR values with a strong correlation of about $R = 0.60$ when tested against an independent dataset. The correlation coefficient drops for both regression algorithms, but more so for the RF. It remains to be seen whether this prediction is good enough to be applied in a practical corpus correction scheme.

The problem that experiment 2 tries to tackle however, is equivalent to confidence measure estimation in ASR in that the same features are used for producing the automatic S&L and a quality estimation of the very same S&L. This might be partly responsible for the worse results (compared to the results for experiment 1).

Experiments 2b and 2c are an extension of experiment 2a. These experiments have shown that it is possible to create models based on resampling strategies that predict values with an error that is more equal across the range and have a smaller prediction error, especially in the bins close to 0. However, this comes with an increase in the prediction error in bins close to 1 and an overall decrease in the correlation coefficient. A problem with the metrics used is that they all prefer models that perform well for densely populated regions (in the current case those with OvR values close to 1).

An interesting point to take into consideration is the large variation in the first bin (0.0 to 0.05). This bin contains two different types of values: a) the predictions that really have been predicted with a value between 0.0 and 0.05 and b) the values that were predicted with an OvR smaller than 0.0 (and were set to 0.0). The dynamic of the features extracted

when segments neither overlap nor touch, might be different than those that result from an overlap from 0 to 0.05. This might be an explanation as to why there is bigger variation than in neighboring bins.

4.7 Conclusion and Future Work

The prediction of transcription word errors as described here seems a promising method making the process of speech corpus annotation more efficient. The presented method based on a SVM will be implemented and made available as a web interface and as a service within the CLARIN infrastructure of the Bavarian Archive for Speech Signals (BAS)[11]. Predicting of the automatic S&L time-alignment errors turns out to be more challenging. Therefore, it remains to be evaluated whether the method can successfully be applied in corpus correction. Some additional features and modifications that could improve MOCCA are listed in the following.

A possible feature, which is inspired by the lattice-based posterior probability approaches (cf. Sec. 4.3.3), would be defined as:

$$f_{nBestRatio} = \frac{\#\text{N-Best Hypothesis}}{\#\text{possible N-Best Hypothesis}} \tag{4.9}$$

where # stands for "number". This would calculate the ratio of the number of alternative hypotheses the decoder outputs, divided by the number of n-best hypotheses that are possible, based on all possible paths through the restricted phonotactic model (cf. Sec. 1.3.3). This ratio can be greater than one, as the number of alternate hypotheses can be high in the lattice due to different starting and end times of segments, compared to the ones that are possible. The normalization of these values based on the lattice of the language model seems like a good opportunity to make these values comparable across utterances and, therefore, add valuable information to the predictor.

A possible feature generally available in ASR systems is a complete language model (usually based on n-grams). With the aid of a language model, it would be possible to

[11] Accessible under `http://hdl.handle.net/11858/00-1779-0000-0028-421B-4`).

check the grammatical correctness of a word, based on its context. This should improve the recognition rate as it is assumed that errors are often ungrammatical (Ghannay et al., 2015).

Another possible feature to detect ungrammatical errors would be to incorporate POS tags of the words that are about to be evaluated. POS tags are available via services (e.g., in Hinrichs et al., 2010), and could therefore be comparatively easily integrated.

An improvement that has nothing to do with features or feature engineering is the oversampling present in experiment 2. It is possible that higher oversampling could be used to balance the prediction accuracy across the range of the OvR even further. Torgo et al. (2013) use oversampling percentages of up to 500%. With higher oversampling in the minority bins, it would be possible to add even more information available for the bins close to 1.

Chapter 5

Summary and Conclusion

5.1 Overall Summary

The studies in the last three chapters have shown that automatic methods can benefit the investigation of regional variation in large speech corpora. Two studies have been conducted relating to the pre-processing of data, one on the validity of an automatically obtained segmentation and labeling (S&L) (cf. Chapter 2) and another on error detection in transcripts and S&L. The latter enables the automatic removal of data that is likely to be wrong (cf. Chapter 4). The third study dealt with the geolocalization of speakers, ultimately based on the speech acoustics of regional variation, and with ways to effectively visualize this variation (cf. Chapter 3).

5.2 The Validity of Automatic Segmentation and Labeling for Duration Studies

In Chapter 2, the validation of an automatic S&L was performed. To do so, a subset was created, consisting of the three dialect groups West Central Bavarian (WCB), East Central Bavarian (ECB), and East Franconian (EF). The latter group was treated as a study-internal reference group, which was reported to behave differently than WCB and ECB. This dataset was then processed by the tool WebMAUS to achieve the alignment

needed for testing. Two experiments were conducted on the resulting data to validate the automatic alignment.

First, a well-known dialect phenomenon – in this case, the Central Bavarian lenition – was investigated based on the automatically processed data (cf. Sec. 2.5). The feature used to demonstrate the validity of the automatic S&L was the phonological dialect feature of the complementary vowel length in Central Bavarian dialects which can be characterized by an acoustic feature called the V/(V+C) ratio (cf. Sec. 2.5). In Standard German, long and short vowels, and lenis and fortis consonants are, in general, freely combinable. Examples of this free combination are, e.g., *Mieder* /miːdɐ/, *Mieter* /miːtɐ/, *Mitte* /mɪtə/, and *Widder* /vɪdɐ/. In Central Bavarian, not all combinations are possible, as short vowels only occur in front of the fortis plosives /p, t, k/ and long vowels only in front of the lenis plosives /b, d, g/.

Evaluating the V/(V+C) ratios in Sec. 2.5 using the automatically obtained S&L data, shows that the results generally comply with descriptions in previous studies. This means, that the complementary vowel length can be observed for ECB speakers, to a lesser extent for WCB speakers, and not at all for EF speakers. This is taken as first evidence that the automatically obtained S&L is a valid basis for evaluation. However, the values for the V/(V+C) ratios were scattered across the range of values, which suggested the existence of noise within the data introduced by the automatic S&L process.

Second, an experiment was conducted to assess the noise emerging from an automatic S&L process and to explicitly compare how a manual correction changes the results in comparison to an automatic one (cf. Sec. 2.6). For this, a subset of the data used in the first study was created, which was manually corrected by a human annotator. Here, it could be seen that the feature values from this subset before manual correction behaved similarly for the presented groups as in the full dataset. After manual correction, the three described dialectal regions became more distinct and the overall range of the values became smaller. Nevertheless, the basic structure of the distributions exhibited the same characteristics before and after correction.

These two experiments shed light on two main aspects. First, the automatic S&L using WebMAUS can be successfully employed for dialectological and variational linguistic

research. This holds true, even if the distributions differ from a manual correction since it was possible to draw the same conclusions from both datasets. Second, young speakers in the *German Today* (GT) corpus still produce the reported dialect feature of complementary vowel length. However, there is evidence that in WCB a sound change is happening, as long vowels can also occur in front of fortis plosives. This category is not present in the speech of ECB speakers.

5.3 Geolocalization of Speaker Origins

In Chapter 3 of this thesis, a geolocalization based on acoustic features from regionally varied speech was investigated. In all experiments in this study, a strict bottom-up approach was pursued. Bottom-up in this case means that the system was not provided with any dialectological or linguistic knowledge. For these experiments, the full set of map task recordings of 641 speakers (328 female, 313 male) from the GT corpus was used, which had already been automatically segmented and labeled.

By pursuing a predictive approach, rather than a descriptive one, the method has an objective measure of success (an improvement over the baseline) and, additionally, has the potential to be applied in, e.g., improving Automatic Speech Recognition (ASR) systems. This could be beneficial as it has been reported that having a model selection positively impacts ASR performance (e.g., Najafian, 2016).

The first experiment aimed to test the feasibility of the approach in a binary decision task using information from only a very short signal part of one phoneme (equivalent to roughly 30 ms − 100 ms). It could be shown that the North/South distinction was easier to perform than the East/West, which is in agreement with the traditional grouping of German dialects. The East/West distinction could be predicted best by the phoneme /øː/, with an accuracy of 0.5791, and the North/South distinction by the best phoneme /z/, with an accuracy of 0.7037. Compared to previous dialect classification attempts this performance, at the first glance, falls short. Nevertheless, the proposed method, based on Random Forests (RFs), has a few major advantages. First, the amount of used information (only one phoneme) to make a prediction is considerably shorter than in earlier approaches.

Second, no dialect corpus with distinct classes was used, but a data-driven separation underlies the prediction. And third, the proposed approach enables to relate the extracted features to be related to dialectological and phonetic variation reported in the literature, something that most other approaches do not allow.

The goal of the second experiment was to investigate a continuous estimation of speaker origins. To my knowledge, this is the first study to examine this kind of geolocalization. It was found that the proposed method predicting a continuous speaker origin using only a small speech sample, as in experiment 1 of this study, improves the prediction over a conservative baseline in an east-west direction by only 6.24% and for north-south by 12.65% for the best performing phoneme /z/. Therefore, it is likely that the proposed approach does not perform well enough to be applicable in ASR model selection. It seems that this can be attributed to several reasons: the limited speech information available, that static features do not capture the nature of speech sounds well enough for a regression task (i.e., a dynamic representation over longer stretches of the complete phoneme segment are necessary), that the intra-speaker variability in producing regional variation conflicts with a reliable prediction based on only a single phoneme, and that the chosen Machine Learning (ML) algorithms are not powerful enough to model the available variation sufficiently. The choice of the ML methods was motivated by an attempt to explain the phonetic variation underlying the prediction. If this requirement is left out, other, presumably more powerful approaches already applied to dialect classification could be considered, such as, e.g., i-vector approaches (e.g., DeMarco et al., 2013) or Deep Neural Networks (DNNs) (e.g., Lopez-Moreno et al., 2014).

After the data from multiple realizations of the phonemes uttered by one speaker were combined, geolocalization performance improved considerably above the proposed conservative baseline. Based on the testing method (Leave-25%-Speaker-out Cross Validation (CV)), every speaker tested was unknown to the system during modeling. This supports the two hypotheses from the last paragraph that intra-speaker variability and the sparse information contained in a single phoneme had an influence on prediction accuracy. Furthermore, this is taken as evidence that regional variation is captured by the recordings in the corpus, that acoustic features alone are able to capture this variation, and that

static features and short-time functionals extracted around the phoneme midpoint carry sufficient information.

Moreover, by training a Decision Tree (DT) using the aggregated data, the division of the geographic space could be explained (in large parts) by already known regional variation and visualized in a way that resembles traditional dialect geography (e.g., Schmidt et al., 2001; Bayerischer Sprachatlas by Hinderling et al., 1996 – 2014) and dialectometric studies (e.g., Goebl, 2010; Nerbonne et al., 2013). The division of the geographic space by the phoneme /z/ in the current study resembles the division reported by König (1989, p. 93–96) strongly, even though, in the current study, the division takes place further south than previously reported. Nevertheless, the separating line between the northern and the southern part of the corpus area coincides closely with the isogloss that separates High and Low German dialects. This is taken as evidence for the validity of this separation, which was generated by a strict bottom-up approach in the current approach.

The current approach differs in three main aspects compared to dialectometric studies, even though both use information theoretic methods for the study of language variation. First, dialectometry is, to a large part and often solely) based on auditory transcripts of regional variation that can be found in atlases: the study performed in Chapter 3 uses only measurable acoustic features. Second, by using a transcript, a mapping of this categorical representation of variation to a *distance measure* is necessary. For example, the Groningen school of dialectometry often does this by using the Levenshtein distance. The already numeric and continuous character of acoustic features renders this unnecessary for the proposed approach, which means that the values can directly be used to visualize and calculate the similarity between sites. Third, the colors used for the visualization of results are different. In the current study, perceptually balanced color gradients (Moreland, 2009) are used to visualize the acoustic feature values on the map. This seems important, as many authors argue that the rainbow colormap is not a good choice for the visualization of continuous variables (Rogowitz et al., 1996; Rogowitz et al., 1998; Borland et al., 2007; Moreland, 2009; Niccoli, 2012). However, using a rainbow colormap also has advantages, as the colors which are drawn from it, which are hard to relate to continuous changes, lead to a more separated-looking map, than does, e.g., a perceptually balanced color scale.

5.4 Confidence Measures in Automatic Segmentation and Labeling

In Chapter 4, an additional pre-processing step called Measure of Confidence for Corpus Analysis (MOCCA) was proposed. This borrows ideas relating to Confidence Measure (CM) in ASR to evaluate the quality of a recognition hypothesis. When creating a speech corpus that can be used for phonetic research, a series of steps have to be performed. Two error-prone steps are orthographic transcription (performed either manually by a human transcriber or automatically by an ASR system) and phonetic S&L. MOCCA attempts to find errors in both. This should lead to an improved quality of the achieved alignment and a reduction of the manual labor required to do so.

The quality estimation of the manually created orthographic transcription was performed at a word-level, which is similar to CMs in ASR regarding granularity. This is because the word-level is often chosen for the decision in ASR CM as well. To make a prediction, the study performed in Chapter 4 evaluates the usefulness of decoder-based features and achieves an accuracy of 78% in distinguishing correctly from incorrectly transcribed words. Based on the distribution that can be assumed for transcriptions errors, i.e., errors that occur only sporadically, this good level of accuracy still leads to many false negatives that have to be checked. Nevertheless, the current method leads to a great reduction in the number of words that have to be manually reviewed, but will also miss around 22% of transcription errors.

Moreover, the algorithms do output a class probability for each word, instead of only a categorial variable (specifying the correctness of a word). Using this probability, it is possible to exclude those cases from the result, in which the classifier was not sure about its own decision. This can be important in cases in which enough data is available for evaluation. Excluding a portion of the data that is likely to be erroneous would further reduce the dataset in a way that benefits the results, as well as it would decrease the amount of manual work involved in the process.

To estimate the quality of the automatically obtained S&L, the Overlap Ratio (OvR) is used, which describes the amount of relative overlap of two segments. In unseen data,

this value can be regarded as the error of a generated S&L. It gives an estimate of the difference between automatically obtained segment boundaries and segment boundaries hypothetically produced by a human labeler. The OvR is predicted by regression algorithms, in the current case RFs and Support Vector Regression (SVR). It can be seen that the correlation between the prediction on an independent test set is strong with $R = 0.60$. This means, even though the same features are used that originally lead to the S&L it is possible to estimate the S&L's quality with them. This circularity is also present in many CMs in ASR.

The prediction accuracy of the OvR across the range could be improved using resampling strategies. In the current study, oversampling was carried out using the Synthetic Minority Over-sampling Technique for Regression (SMOTER). By applying both over- and undersampling, the final model resulted in a more homoscedastic prediction error over the range of values.

This extra effort to equalize the error across the range of values was necessary, as it is important to predict not only values close to a good overlap reliably, but also, and maybe even more importantly, those values close to a missing overlap.

5.5 Conclusion

It has been demonstrated that automatic methods can be applied to benefit the research of regional language variation. The performed validation in Chapter 2 and the spotting of erroneous transcripts and falsely set segment boundaries in Chapter 4 are valuable preprocessing steps for corpus creation. The approach in Chapter 3 allows the description and modeling of regional variation without manual labor.

The findings from this thesis could lead to the use of more acoustic features from large speech corpora in research of regional variation, as even large speech corpora can be processed and then used by individual researchers. That is, because the huge effort needed for a phonetic annotation is reduced by the need to only create a, much more manageable, orthographic transcription and the S&L is executed automatically. Performing an orthographic transcription is still a time-consuming task. However, to efficiently work

with signal files, an orthographic transcription of a part or the complete signal is often performed. This is done, as otherwise it becomes very cumbersome to find the correct files. In those cases in which an orthographic transcript already exists, the corpus at hand can be enriched by an automatic phonetic S&L, with only little additional human effort. It has further been shown that this automatic S&L is sufficient for analyzing the corpus, even though the orthographic transcription used as input does often not reflect regional variants and the automatic S&L process leads to more noise than a manual S&L process. However, by using more data, the conclusions that can be drawn remain the same, and an existing automatic S&L can later still be refined if deemed necessary and affordable.

The method proposed in Chapter 3 takes this one step further and uses ML techniques to model regional variation with the goal of predicting a speaker's origin. Depending on the algorithms, this has the advantage that the model trained using data containing regional variation can be used to describe a dialect, but also to predict a speaker origin. It has been shown that predictive models, in combination with features that can be extracted automatically from the speech signal, are suitable to generate insights into the given data.

By employing a DT, the connection between the applied acoustic features and the geographic space was unveiled and visualized on a map. These visualizations resemble the output of dialectometric studies regarding connection of variation and geographic space. Taking acoustic features into consideration might allow the re-evaluation of already existing dialectal knowledge based on a manual, auditory transcript by a method that is more objective, data-driven, and bottom-up. The fact that the approach proposed in this thesis and traditional dialect geography produce similar distributions not only on different datasets, but also employ inherently different methods of capturing the variation present in speech, is taken as a mutual validation of both approaches with respect to each other.

Appendix A

First Appendix

A.1 Bands of Semi-Tone Spectrum (STS) feature

#	frequency	#	frequency	#	frequency	#	frequency
0	55.000	25	233.082	49	932.328	73	3729.310
1	58.270	26	246.942	50	987.767	74	3951.066
2	61.735	27	261.626	51	1046.502	75	4186.009
3	65.406	28	277.183	52	1108.731	76	4434.922
4	69.296	29	293.665	53	1174.659	77	4698.636
5	73.416	30	311.127	54	1244.508	78	4978.032
6	77.782	31	329.628	55	1318.510	79	5274.041
7	82.407	32	349.228	56	1396.913	80	5587.652
8	87.307	33	369.994	57	1479.978	81	5919.911
9	92.499	34	391.995	58	1567.982	82	6271.927
10	97.999	35	415.305	59	1661.219	83	6644.875
11	103.826	36	440.000	60	1760.000	84	7040.000
12	110.000	37	466.164	61	1864.655	85	7458.620
13	116.541	38	493.883	62	1975.533	86	7902.133
14	123.471	39	523.251	63	2093.005	87	8372.018
15	130.813	40	554.365	64	2217.461	88	8869.844
16	138.591	41	587.330	65	2349.318	89	9397.273
17	146.832	42	622.254	66	2489.016	90	9956.063
18	155.563	43	659.255	67	2637.020	91	10548.082
19	164.814	44	698.456	68	2793.826	92	11175.303
20	174.614	45	739.989	69	2959.955	93	11839.822
21	184.997	46	783.991	70	3135.963	94	12543.854
22	195.998	47	830.609	71	3322.438	95	13289.750
23	207.652	48	880.000	72	3520.000	96	14080.000
24	220.000						

Table A.1: The indices and frequencies (in Hz) of the STS features (index of feature denoted by '#'). The frequencies are calculated with the formula $\sqrt[12]{2}^{(n-36)} \cdot 440\,\mathrm{Hz}$, where n denotes the feature index.

A.2 Boxplots of Feature Values

A.2.1 Binary North/South Classification

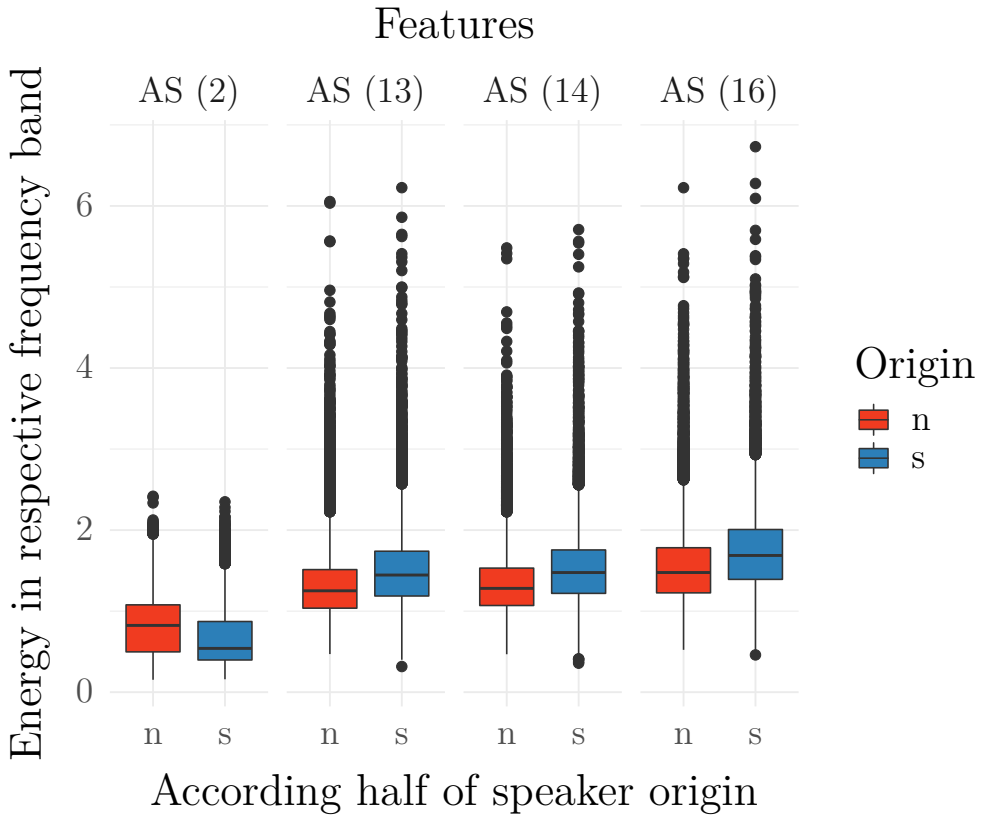


Figure A.1: Boxplots of the feature value AS (2), AS (13), AS (14), and AS (16) of 46,566 produced /z/ from the GT corpus. Groups 'n' (North) and 's' (South) are based on the North/South separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.4.

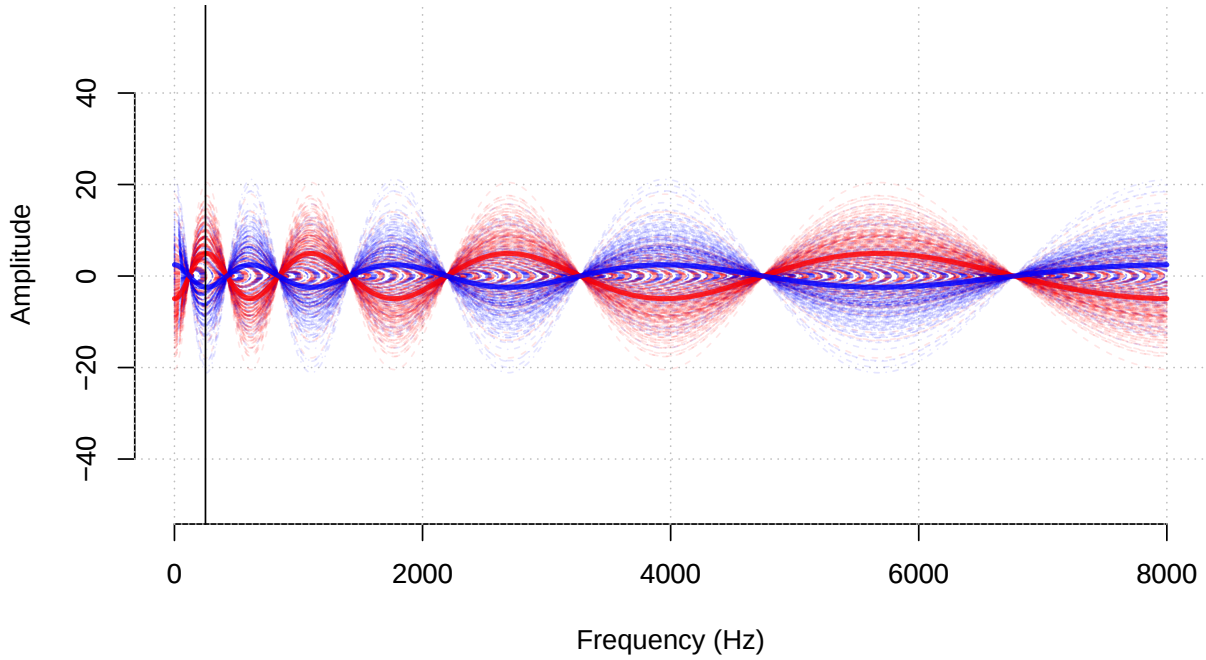


Figure A.2: Resynthesis of feature values MFCC (8) of 46,566 produced /z/ from the GT corpus. Each line corresponds to the resynthesis of the average of a speaker's MFCC coefficients. Lines plotted in red belong to the North group and lines plotted in in blue to the South group based on the North/South separating line defined in Sec. 3.8.2. The two thick lines represent the resynthesis of the averaged values for each group. The vertical line plotted at 250 Hz indicates the center of a possible voice bar. For more information cf. Sec. 3.8.4.

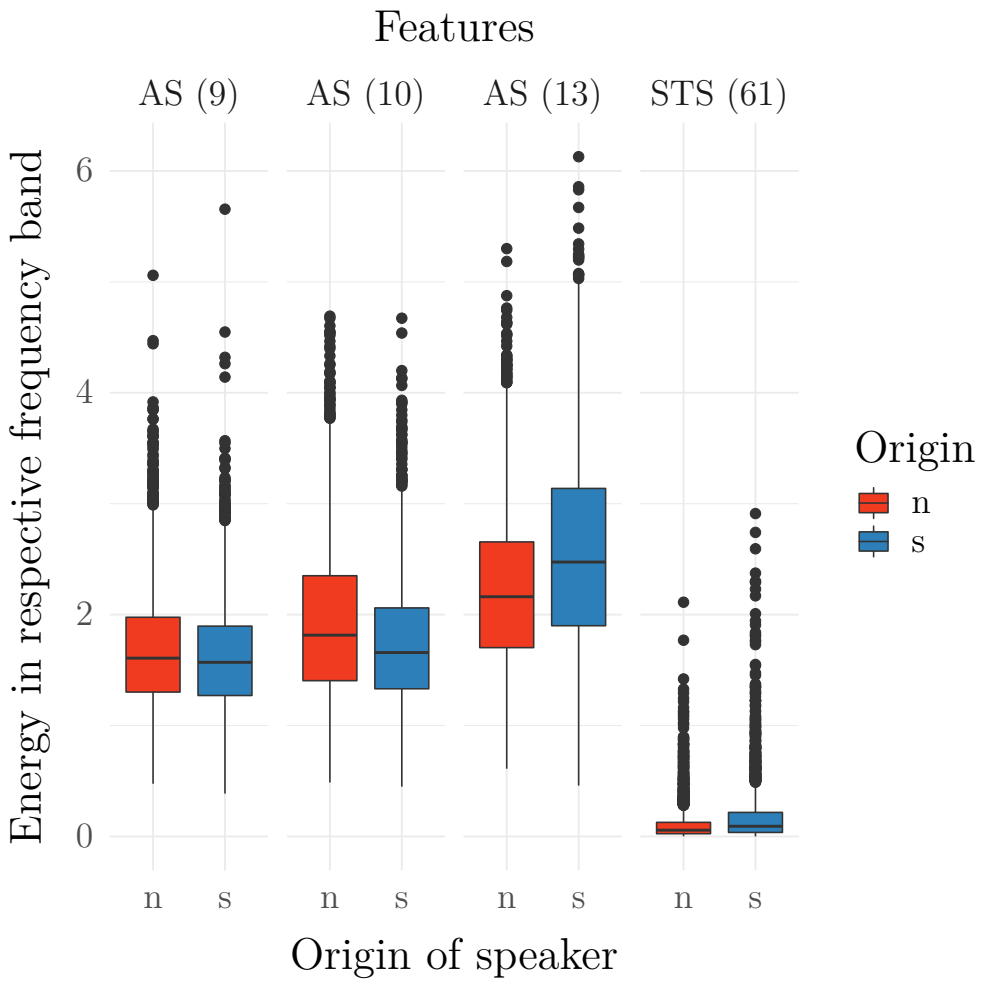


Figure A.3: Boxplots of the feature value AS (9), AS (10), AS (13), and STS (61) of 5764 produced /øː/ from the GT corpus. Groups 'n' (North) and 's' (South) are based on the North/South separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.4.

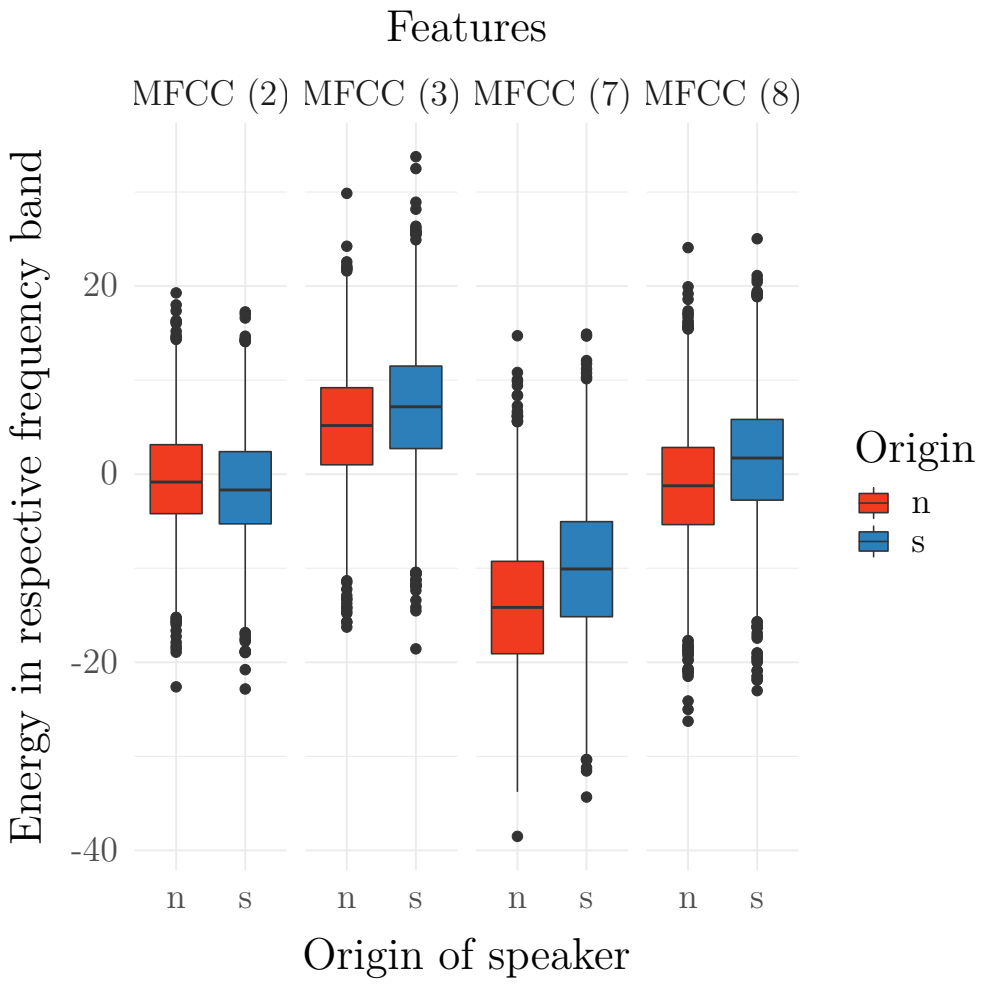


Figure A.4: Boxplots of the feature value MFCC (2), MFCC (3), MFCC (7), and MFCC (8) of 5764 produced /øː/ from the GT corpus. Groups 'n' (North) and 's' (South) are based on the North/South separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.4.

A.2.2 Binary East/West Classification

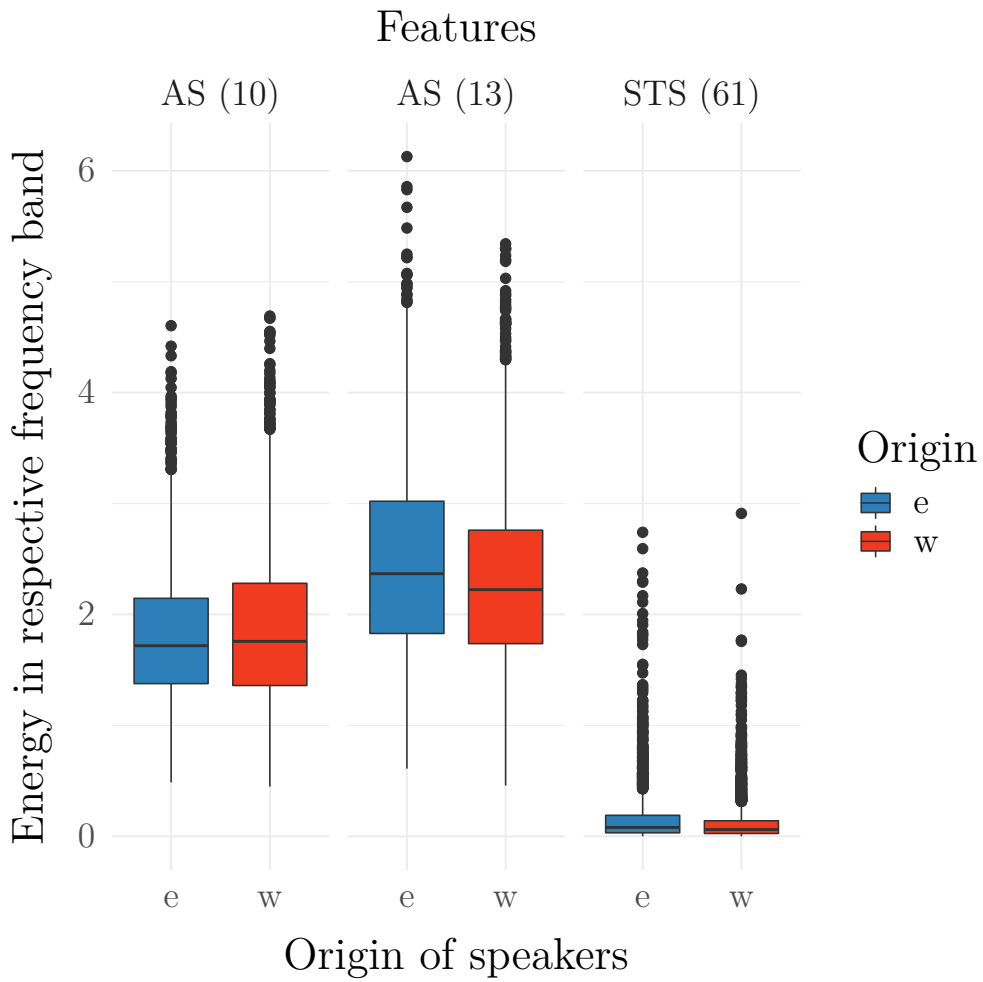


Figure A.5: Boxplots of the feature values AS (10), AS (13), and STS (61) of 5764 produced /øː/ from the GT corpus. Groups 'e' (East) and 'w' (West) are based on the East/West separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.5.

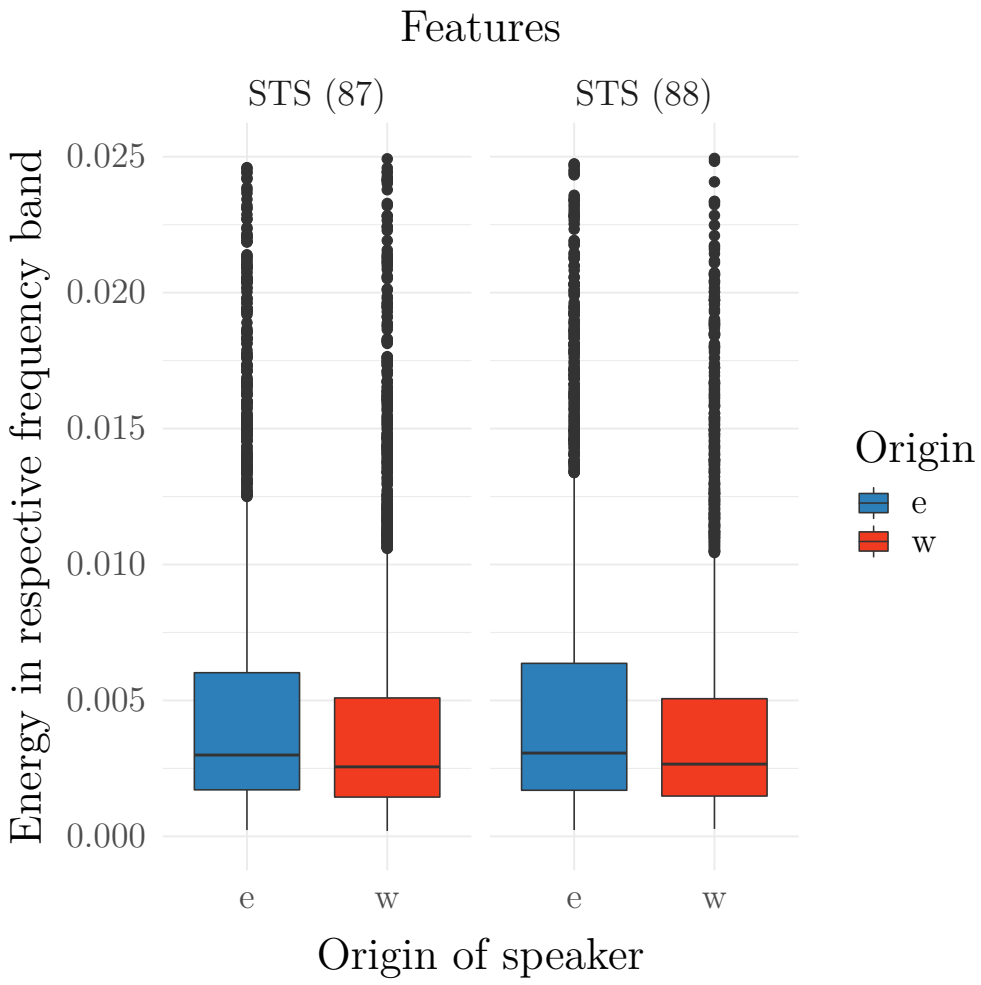


Figure A.6: Boxplots of the feature values of STS (87), and STS (88) of 5531 produced /øː/ from the GT corpus. Groups 'e' (East) and 'w' (West) are based on the East/West separating line defined in Sec. 3.8.2. The y-axis is manually limited to a range between 0 and 0.025 to allow for a better spread of the quantiles (this means that 233 of the furthest outliers are not shown). For more information cf. Sec. 3.8.5.

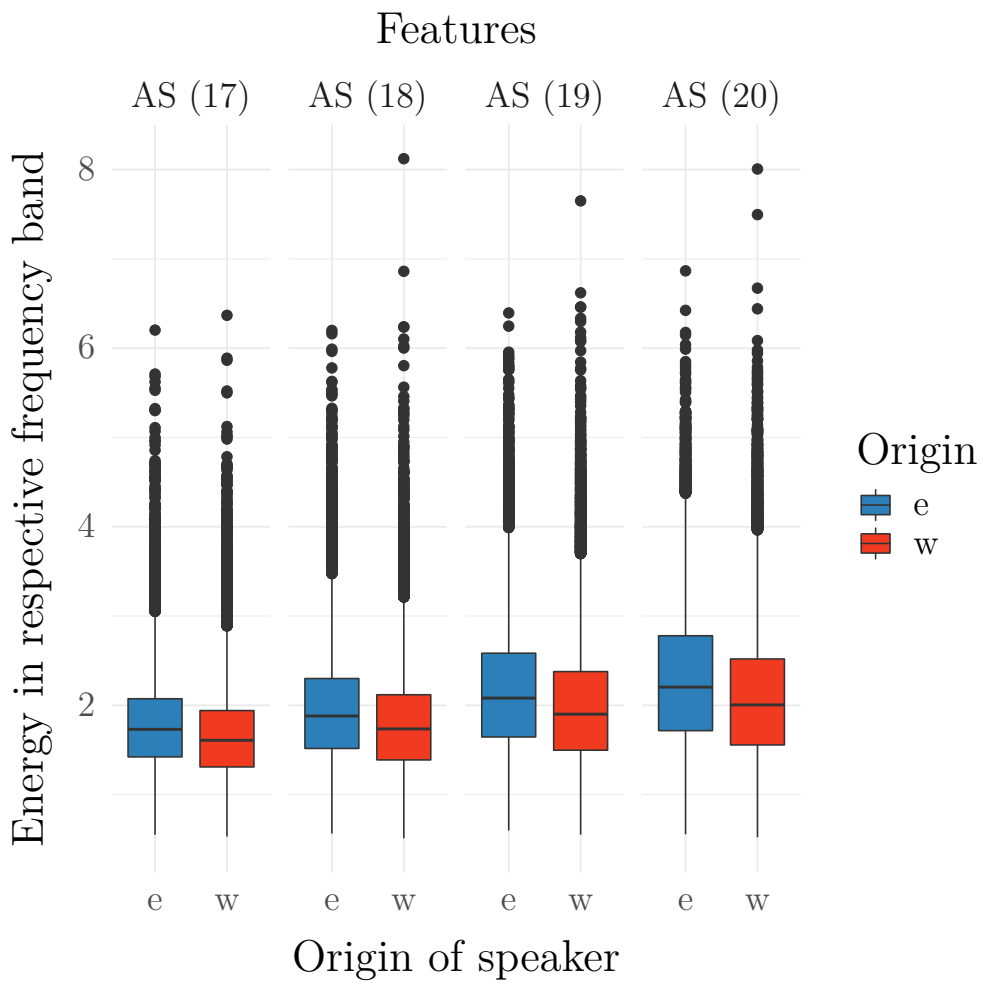


Figure A.7: Boxplots of the feature values for AS (17), AS (18), AS (19), and AS (20) of 46,566 produced /z/ from the GT corpus. Groups 'e' (East) and 'w' (West) are based on the East/West separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.5.

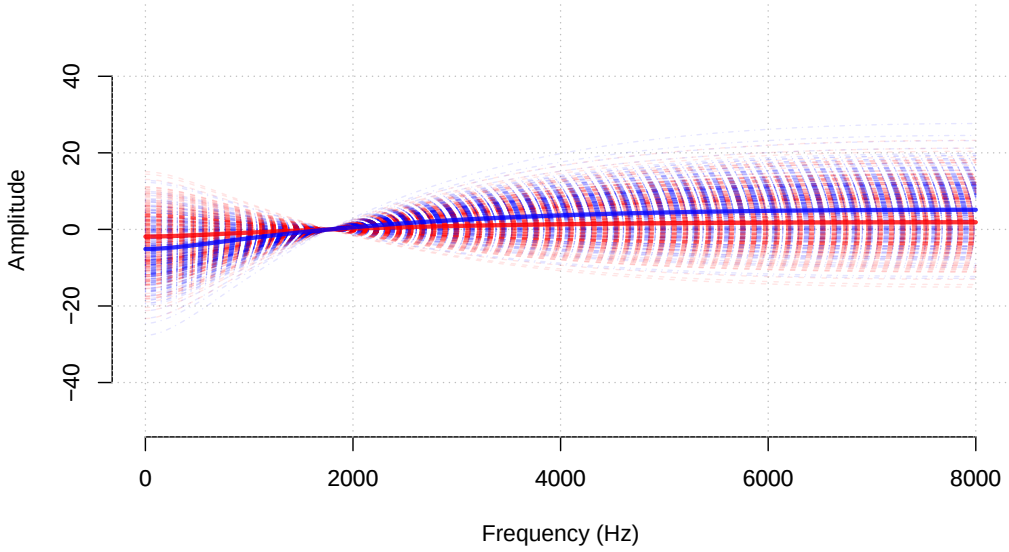


Figure A.8: Resynthesis of the feature value MFCC (1) of 46,566 produced /z/ from the GT corpus. Each line corresponds to the resynthesis of the average of a speakers MFCC coefficients. Lines plotted in red belong to the North group and lines plotted in blue to the South group, based on the North/South separating line defined in Sec. 3.8.2. The two thick lines represent the resynthesis of the averaged values for each group. For more information cf. Sec. 3.8.5.

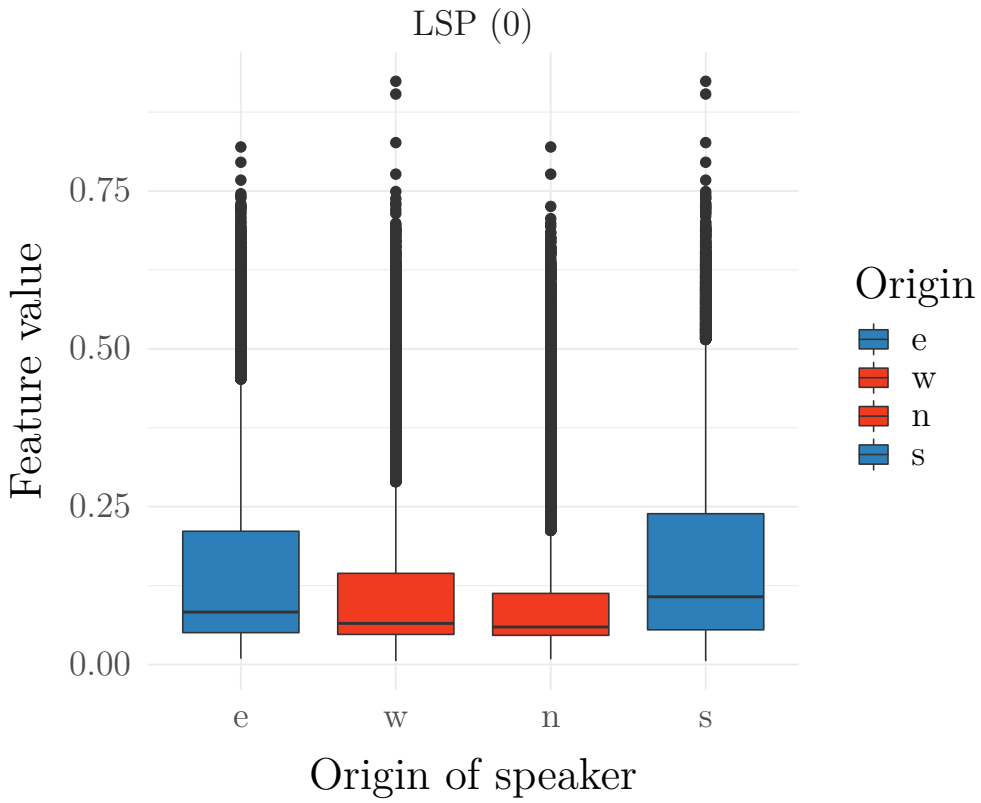


Figure A.9: Boxplots of the feature value LSP (0) of 46,566 produced /z/ from the GT corpus. Groups 'e' (East), 'w' (West), 'n' (North), and 's' (South) are based on the East/West and North/South separating line defined in Sec. 3.8.2. For more information cf. Sec. 3.8.5.

A.2.3 Binary Classification Variable Importance (VI) Comparison

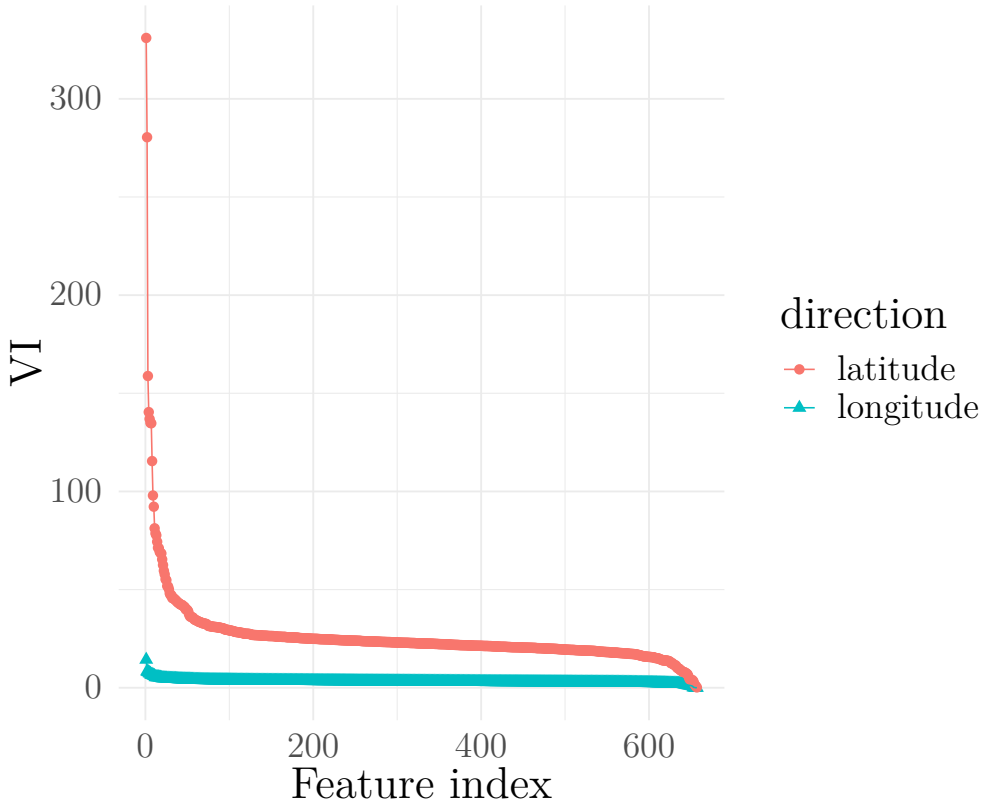


Figure A.10: Scatterplot of the VI of the best performing phonemes for each direction for non-zero VIs. Contrary to all other results reported in Sec. 3.8, the VI is reported for $mtry = 100$. This is necessary as *mtry* influences the model's complexity, and otherwise, the values for the VI are not comparable. For more information cf. Sec. 3.8.7.

A.2.4 Regression North-South Dimension

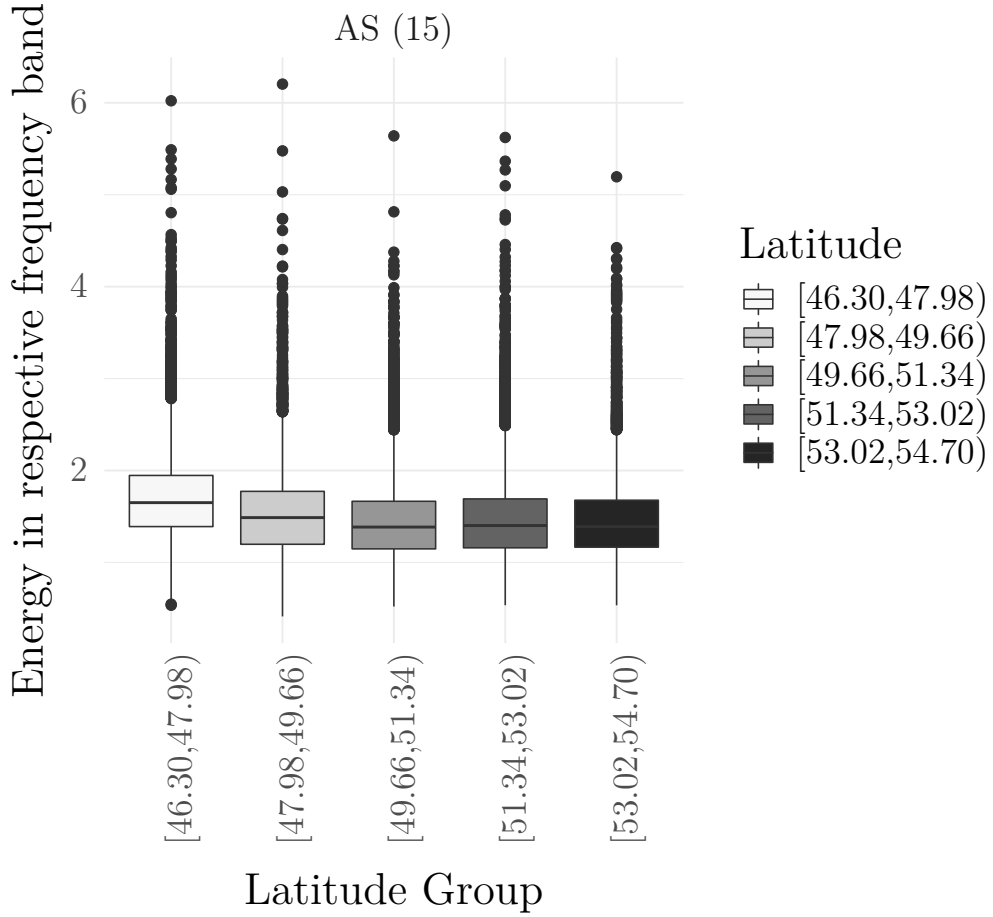


Figure A.11: Boxplots of the feature value AS (15) of 46,566 produced /z/ from the GT corpus. Latitude is binned in 5 equal spaced intervals between 46.30° (most south) and 54.70° (most north). For more information cf. Sec. 3.9.4.

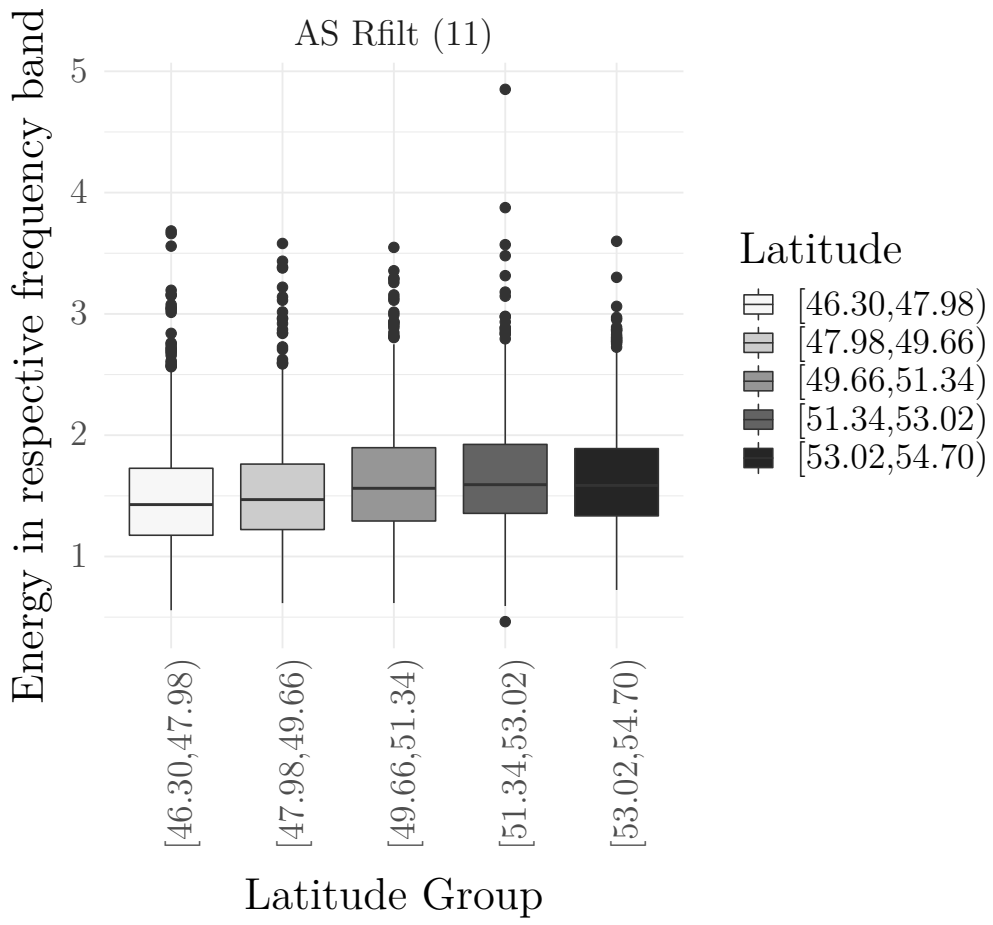


Figure A.12: Boxplots of the feature value AS Rfilt (11) of 5764 produced /ø:/ from the GT corpus. Latitude is binned in 5 equal spaced intervals between 46.30° (most south) and 54.70° (most north) For more information cf. Sec. 3.9.4.

A.2.5 Regression East-West Dimension

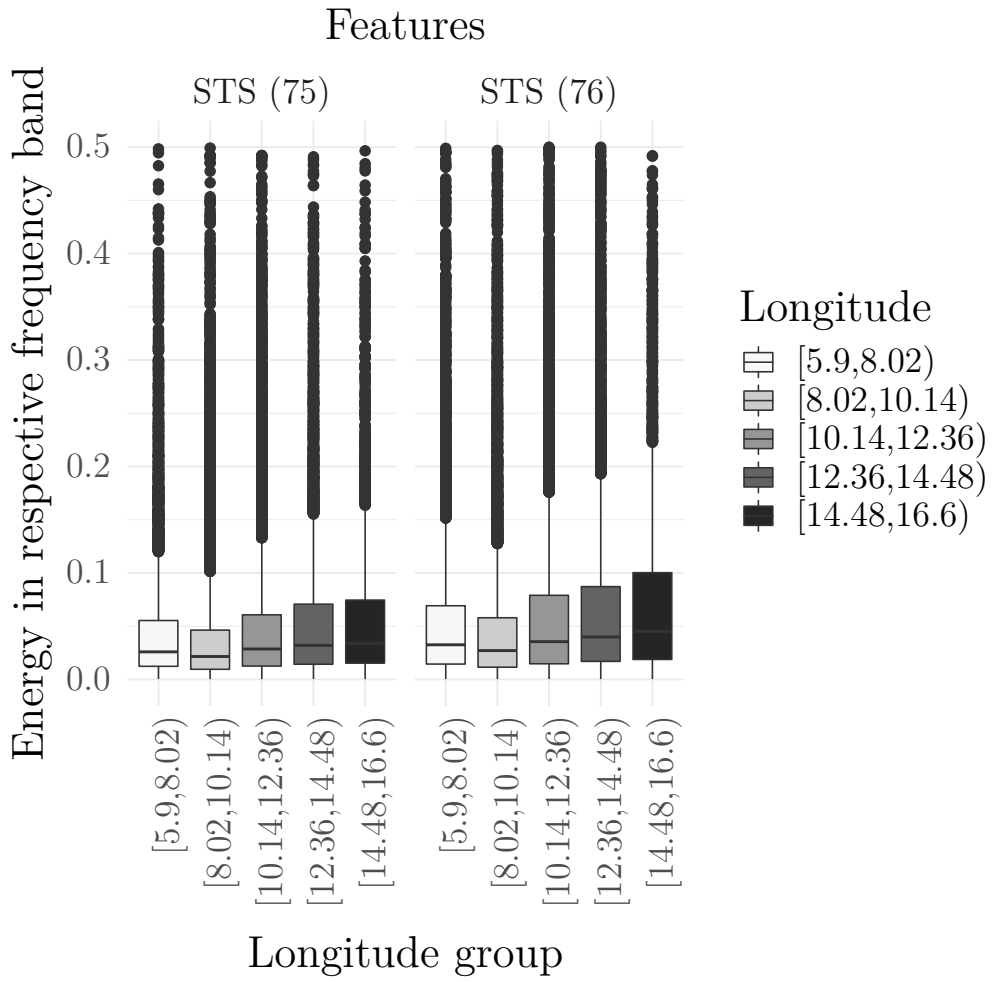


Figure A.13: Boxplots of the feature value STS (75) and STS (76) of 46,566 produced /z/ from the GT corpus. Longitude is binned in 5 equal spaced intervals between 5.9° (westmost interval) and 16.6° (eastmost interval) The y-axis is manually limited to a range between 0 and 0.5 to allow for a better spread of the quantiles (this means that 552 of the furthest outliers are not shown). For more information cf. Sec. 3.9.5.

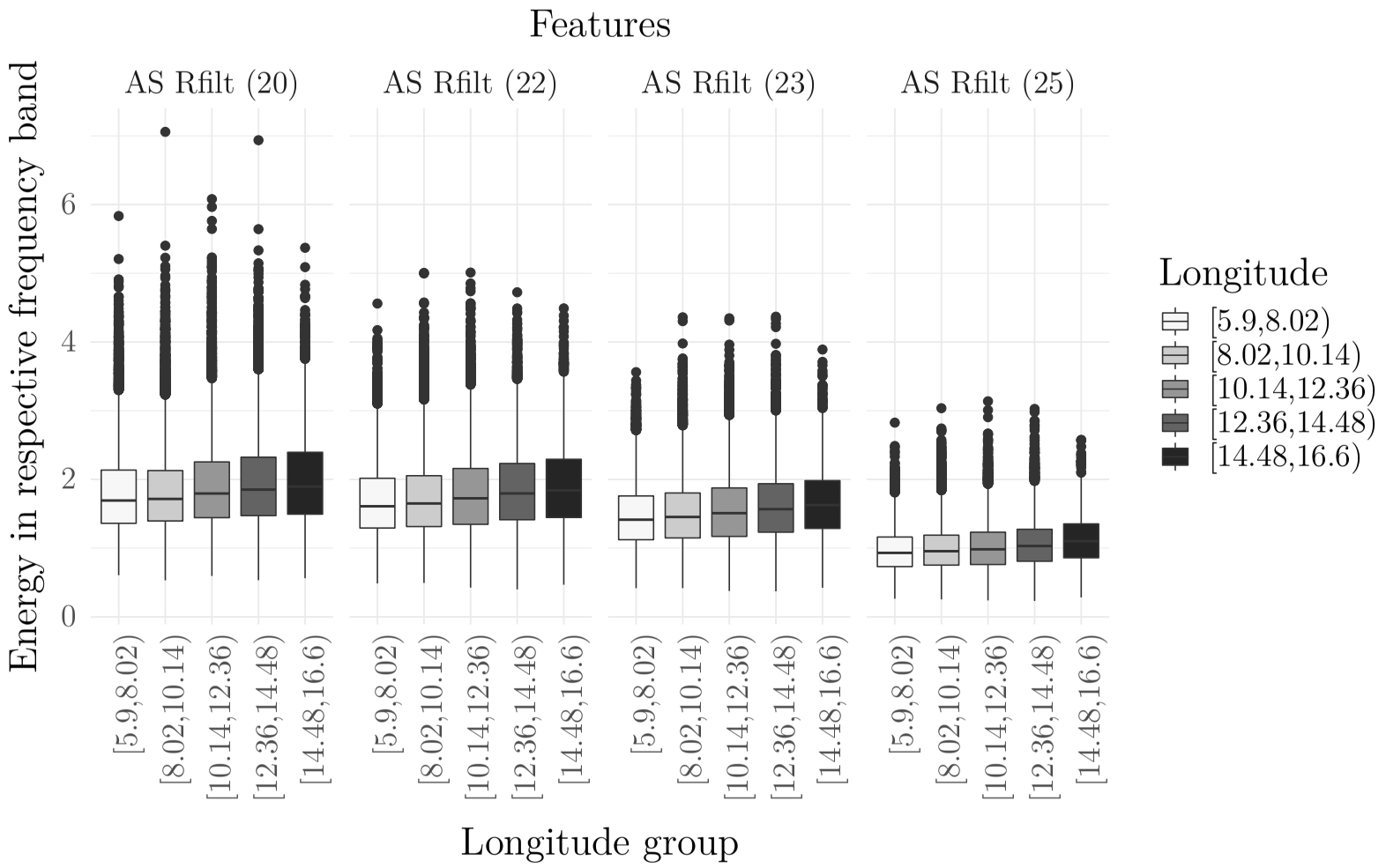


Figure A.14: Boxplots of the feature value AS Rfilt (20), AS Rfilt (22), AS Rfilt (23), and AS Rfilt (25), of 46,566 produced /z/ from the GT corpus. Longitude is binned in 5 equal spaced intervals between 5.9° (westmost interval) and 16.6° (eastmost interval). For more information cf. Sec. 3.9.5.

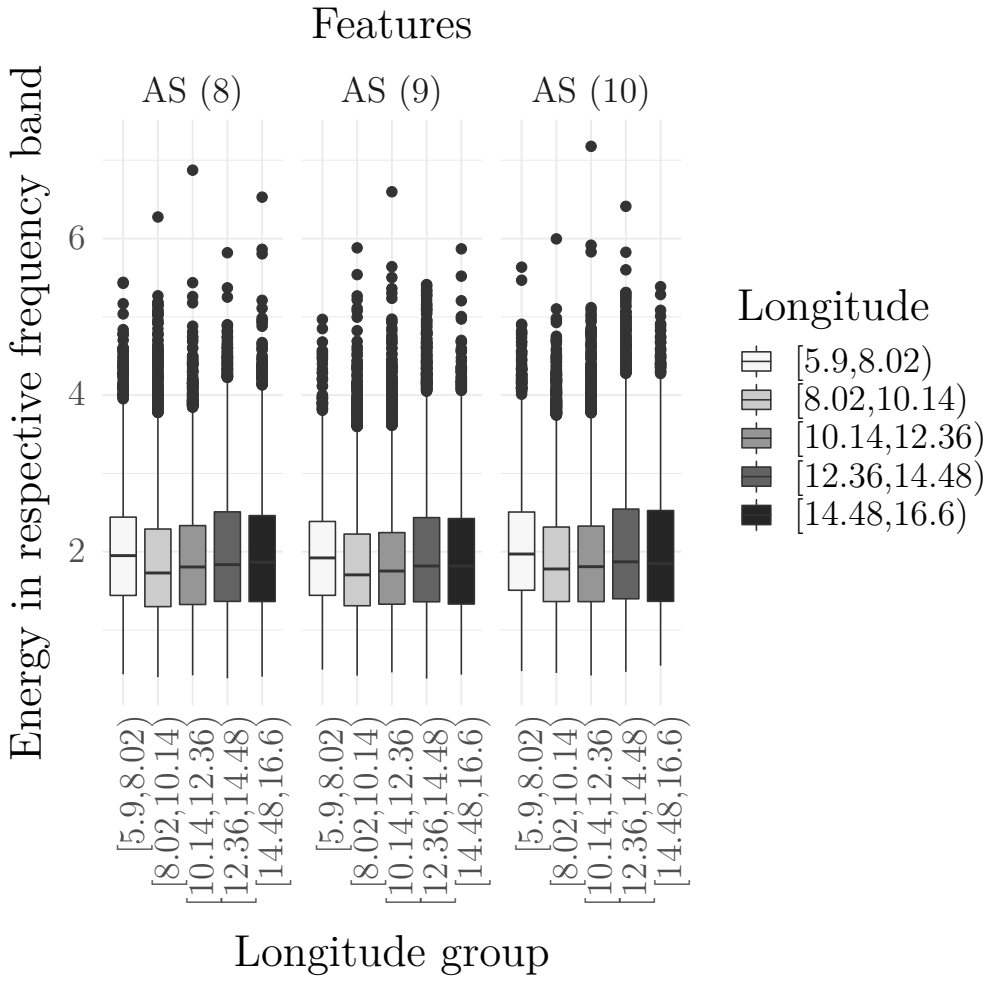


Figure A.15: Boxplots of the feature value AS (8), AS (9), and AS (10), of 28,693 produced /ɛː/ from the GT corpus. Longitude is binned in 5 equal spaced intervals between 5.9° (westmost interval) and 16.6° (most East). For more information cf. Sec. 3.9.5.

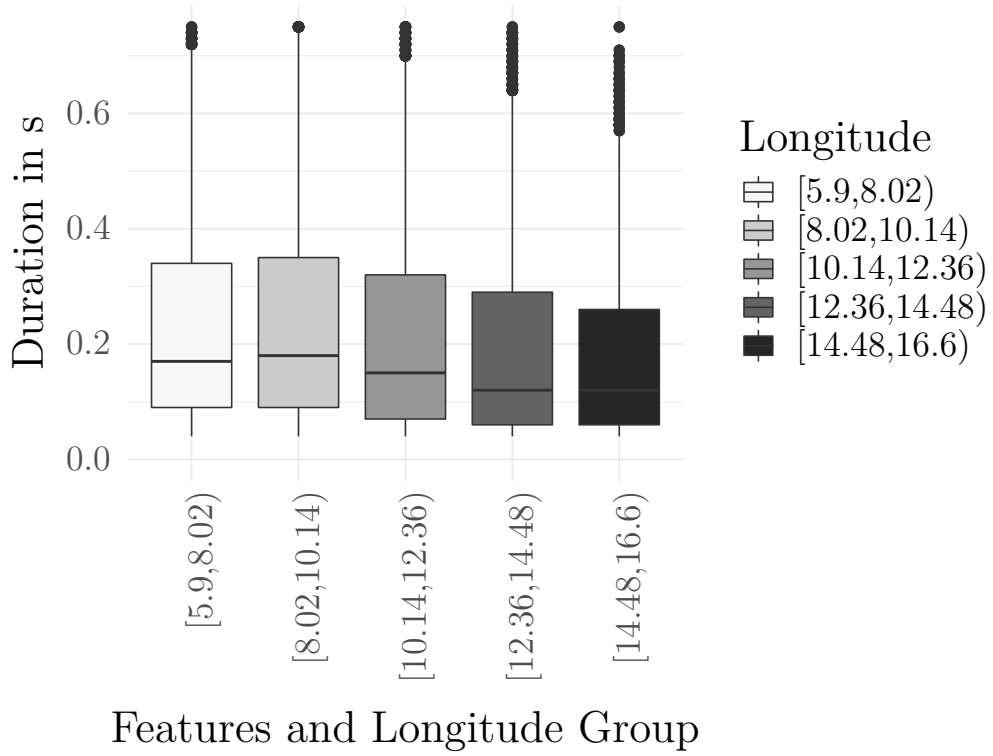


Figure A.16: Boxplots of values of feature duration of 28,693 produced /ɛː/ from the GT corpus. Longitude is binned in 5 equal spaced intervals between 5.9° (westmost interval) and 16.6° (eastmost interval). For more information cf. Sec. 3.9.5.

A.2.6 Phonetic Interpretation of the Decision Trees

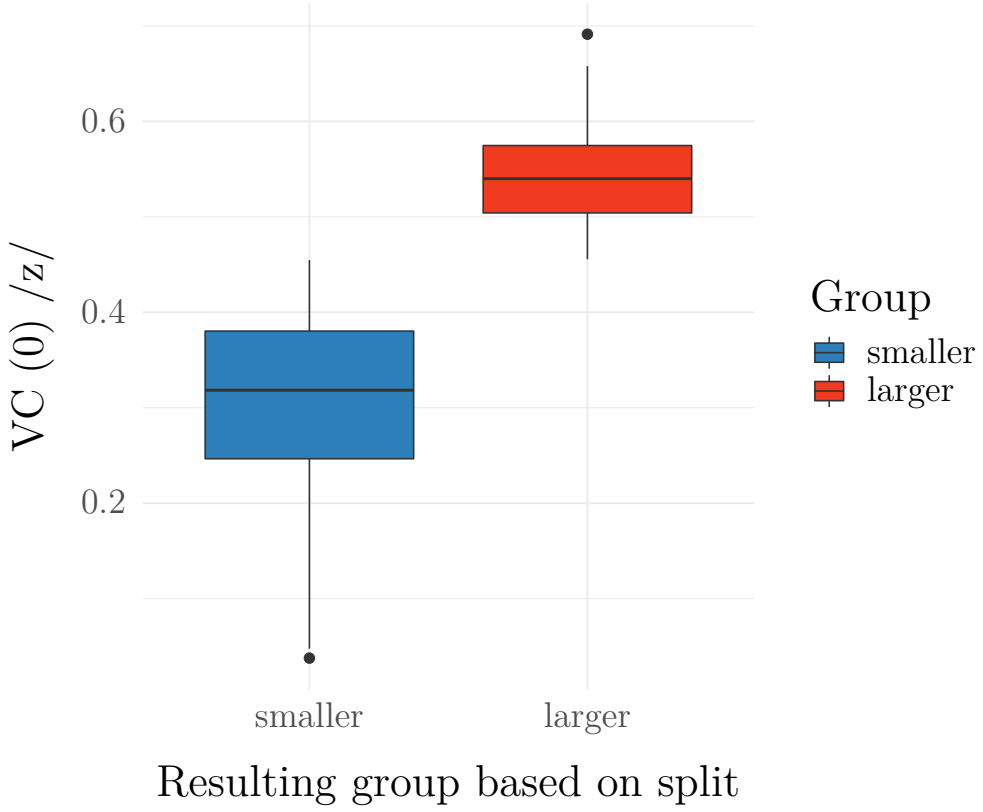


Figure A.17: Boxplots of the feature value VC (0) for /z/ for 641 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), speakers belong to either the blue (< 0.4550677) or the red group (≥ 0.4550677).

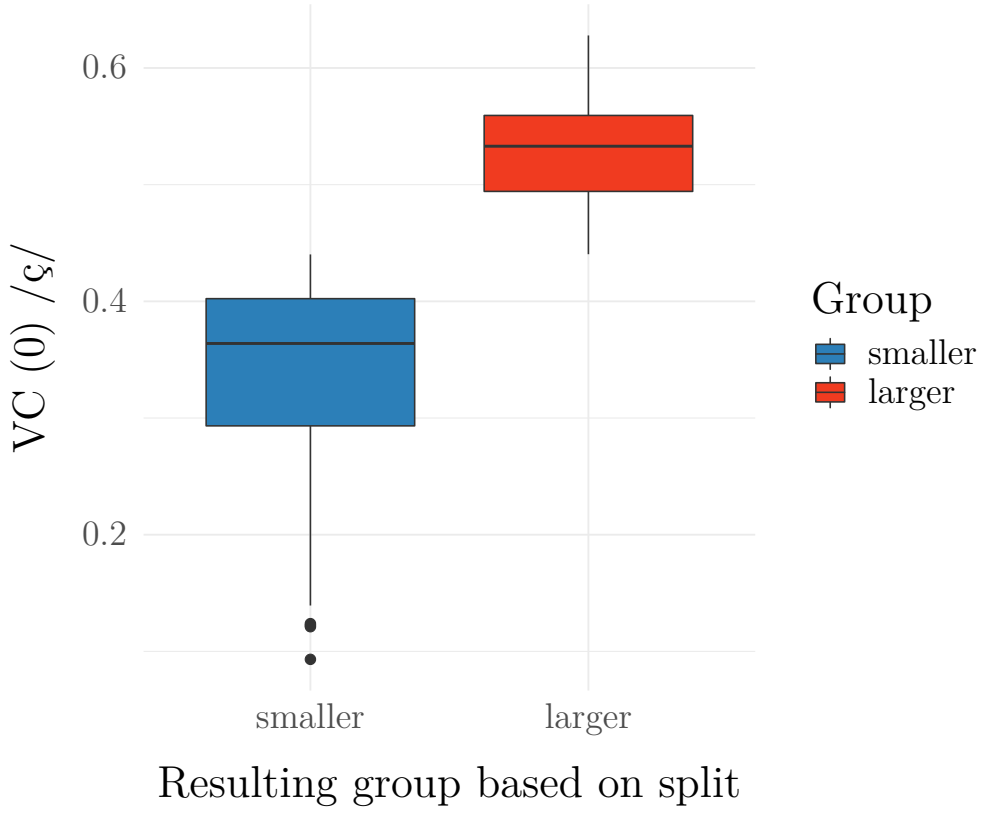


Figure A.18: Boxplots of the feature value VC (0) for /ç/ for 326 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< 0.4402981) or the red group (≥ 0.4402981).

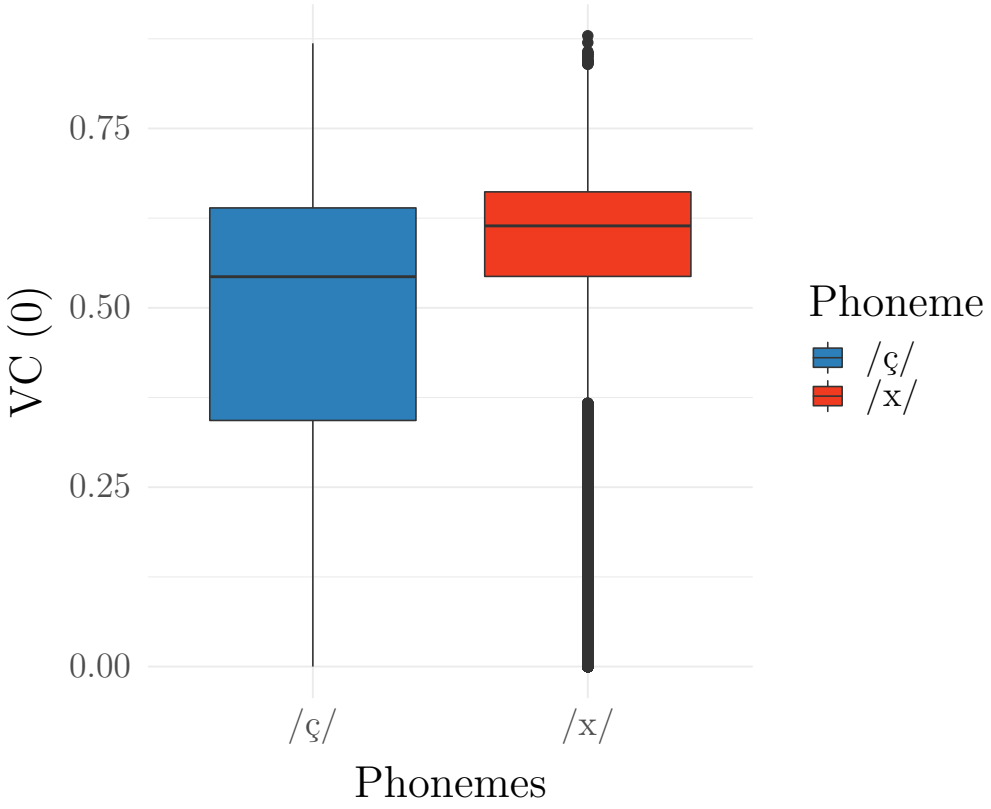


Figure A.19: Boxplots of the feature value VC (0) for /ç/ (49,738 realizations) and /x/ (31,084 realizations) for all 641 speakers.

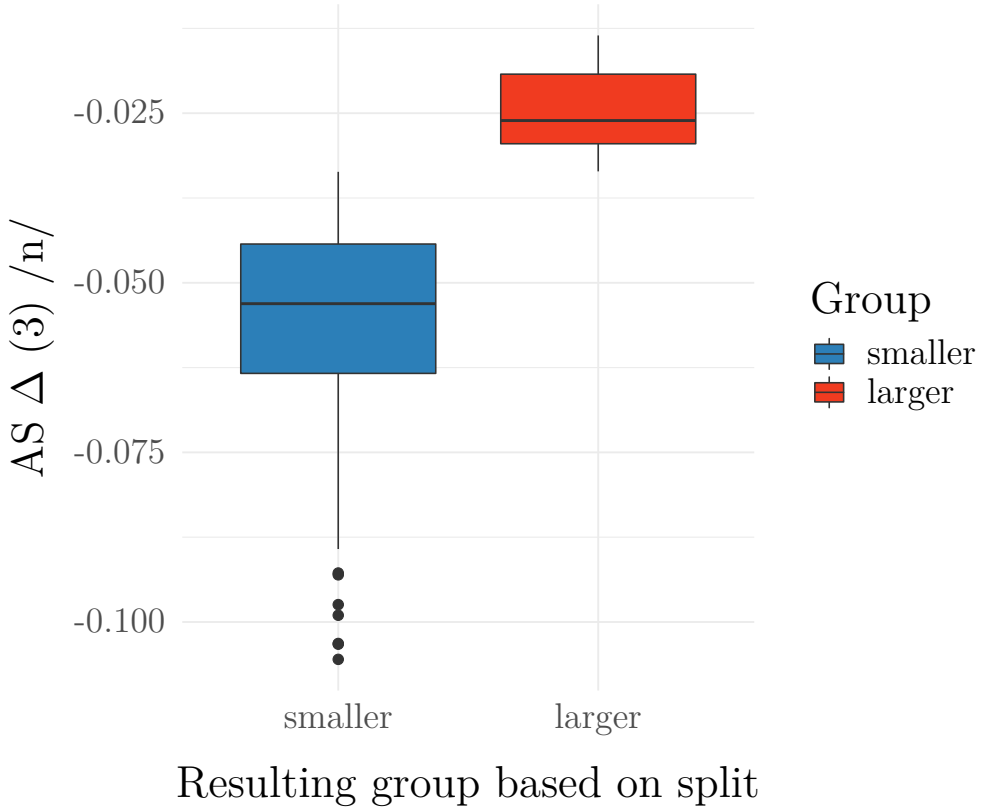


Figure A.20: Boxplots of the feature value AS Δ (3) for /n/ for 231 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< -0.03360892) or the red group (≥ -0.03360892).

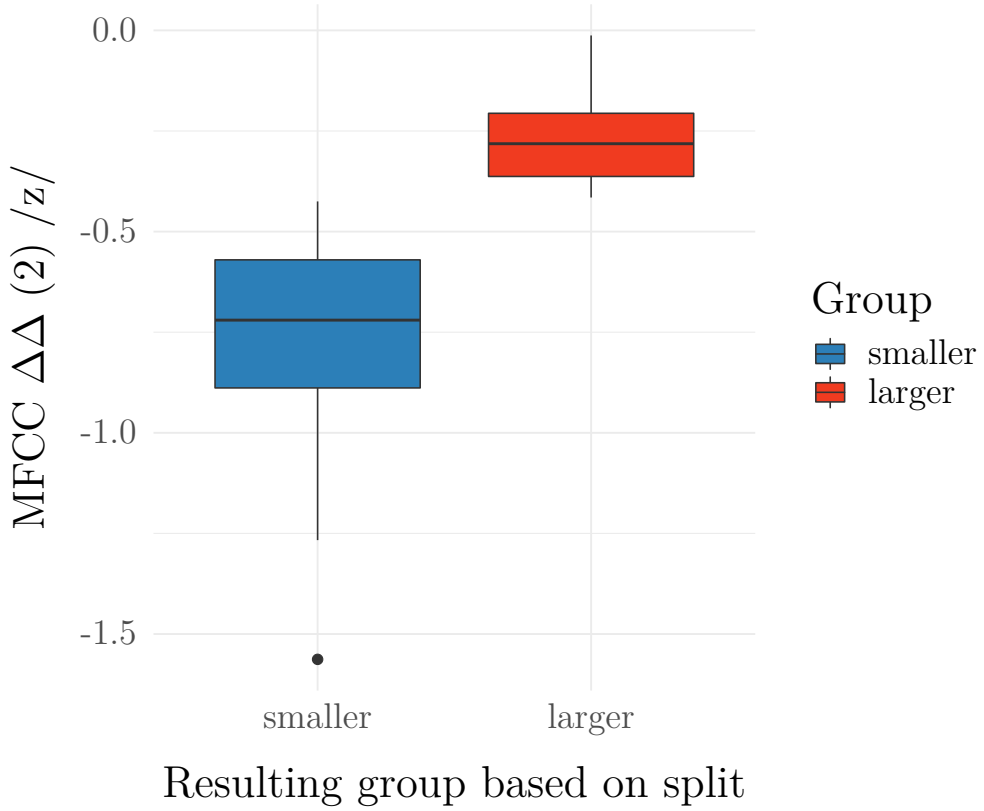


Figure A.21: Boxplots of the feature value MFCC ΔΔ (2) for /z/ for 95 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< -0.4201756) or the red group (≥ -0.4201756).

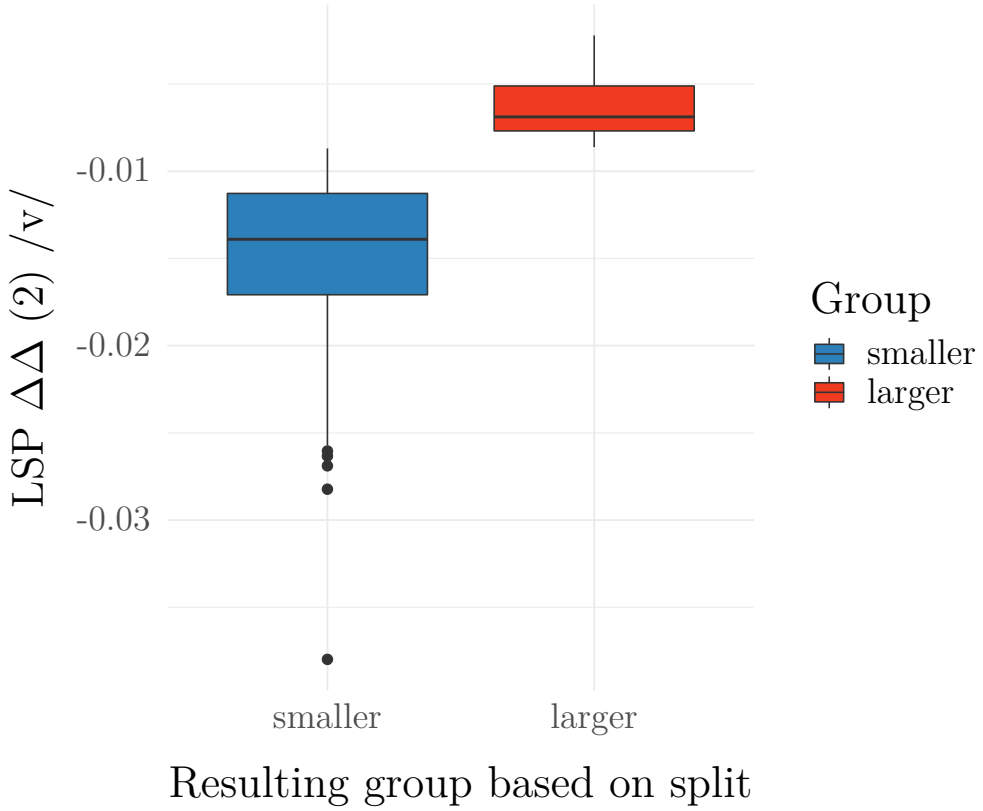


Figure A.22: Boxplots of the feature value LSP ΔΔ (2) for /v/ for 315 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< -0.008655971) or the red group (≥ -0.008655971).

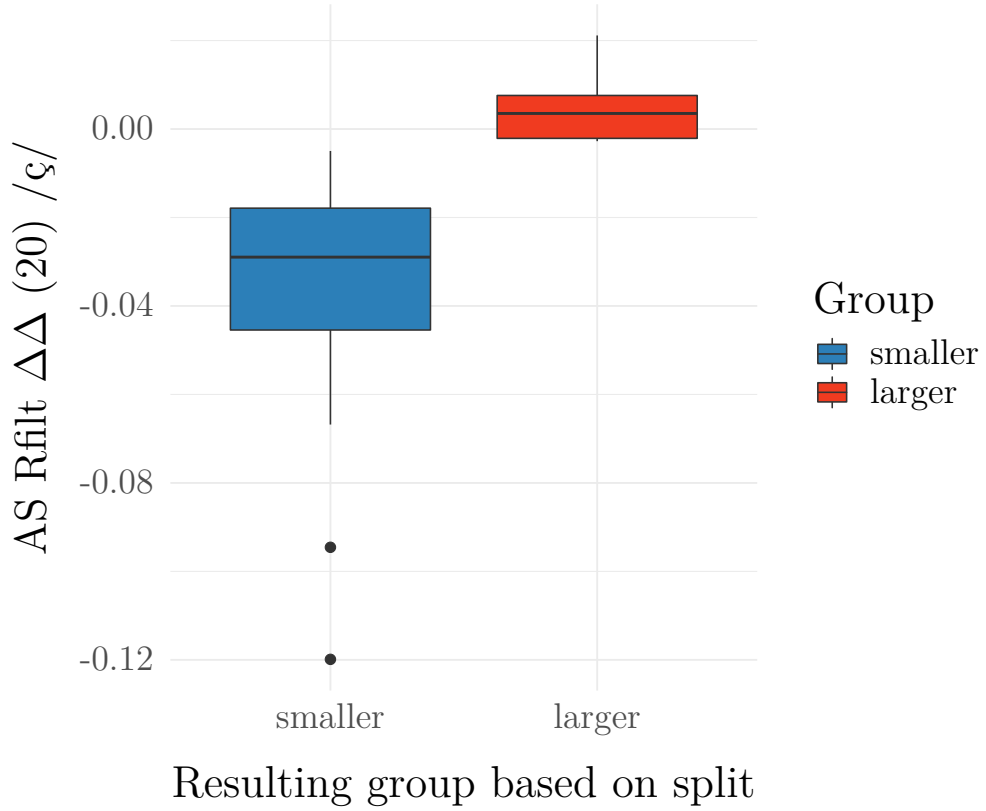


Figure A.23: Boxplots of the feature value AS Rfilt ΔΔ (20) for /ç/ for 58 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< -0.003853369) or the red group (≥ -0.003853369).

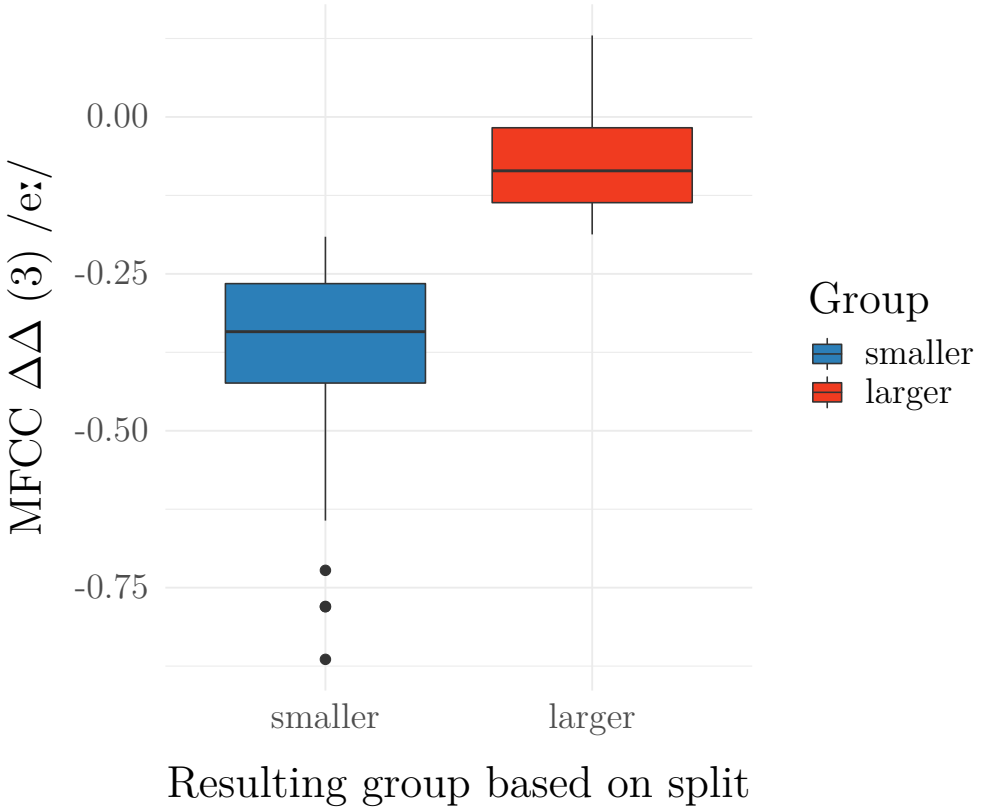


Figure A.24: Boxplots of the feature value MFCC ΔΔ (3) for /eː/ for 257 speakers. Based on the value selected for the according split in the DT (cf. Fig. 3.14), the speakers belong to either the blue (< -0.1889846) or the red group (≥ -0.1889846).

A.3 Estimation of Speaker position - Experiment 3 - Decision Tree Model

The following shows a textual representation of the model of the decision tree based on the reduced feature set from Sec. 3.10.4 as output by the R Programming Language (R) implementation. Each line represents a node, using the following format:

```
node), split, n, deviance, yval
      * denotes terminal node
```

in which *node* is the node number, *split* is the variable[1] used for splitting and the value at which the split is performed, *n* is the number of speakers in that group, *deviance* an estimation of the goodness of fit (for more detail cf. Therneau et al., 2018) and *yval* the average of all elements in this node/leaf.

```
 1) root 641 3141.75500 50.01710
   2) candVoicing.0._z< 0.4550677 326  842.22830 48.62966
     4) candVoicing.0._C>=0.4402981 231  400.43640 48.08485
       8) audSpec_de.3._n< -0.03360892 184  218.17590 47.78054
        16) audSpec_Rfilt_de_de.11._I>=-0.03390794 142   85.02359 47.50972 *
        17) audSpec_Rfilt_de_de.11._I< -0.03390794 42   87.52399 48.69619 *
       9) audSpec_de.3._n>=-0.03360892 47   98.51711 49.27617 *
     5) candVoicing.0._C< 0.4402981 95  206.50230 49.95442
      10) pcm_fftMag_mfcc_de_de.2._z>=-0.4201756 31   21.47871 48.84645 *
      11) pcm_fftMag_mfcc_de_de.2._z< -0.4201756 64  128.53500 50.49109
        22) audSpec_Rfilt.7._E_t>=1.271268 46   55.44759 50.02152 *
        23) audSpec_Rfilt.7._E_t< 1.271268 18   37.02378 51.69111 *
   3) candVoicing.0._z>=0.4550677 315 1022.52700 51.45298
     6) lspFreq_de_de.2._v>=-0.008655971 58  132.65730 49.74052
      12) audSpec_Rfilt_de_de.20._C>=-0.003853369 16   17.19759 48.04562 *
      13) audSpec_Rfilt_de_de.20._C< -0.003853369 42   51.98759 50.38619 *
     7) lspFreq_de_de.2._v< -0.008655971 257  681.39690 51.83946
      14) pcm_fftMag_mfcc_de_de.3._e_t>=-0.1889846 76  176.13310 50.97895
```

[1] With the structure being the feature as output by openSMILE, followed by an underscore and the phoneme in Speech Assessment Methods Phonetic Alphabet (SAM-PA).

```
    28) lspFreq_de_de.1._v>=-0.003868417 18    38.23025 49.67167 *
    29) lspFreq_de_de.1._v< -0.003868417 58    97.59444 51.38466 *
15) pcm_fftMag_mfcc_de_de.3._e_t< -0.1889846 181   425.35810 52.20077
    30) lpcCoeff.1._C>=1.629062 61   134.08960 51.48639
      60) pcm_fftMag_mfcc.1._s>=-6.37478 19    25.31127 50.40526 *
      61) pcm_fftMag_mfcc.1._s< -6.37478 42    76.52384 51.97548 *
    31) lpcCoeff.1._C< 1.629062 120   244.31310 52.56392 *
```

A.4 Estimation of Speaker position - Experiment 3 - Split Models for North and South Half

A.4.1 Differences in Experimental Design to Original Experiment

The experiment in 3.10 was partially repeated due to bad performance for the longitudinal direction. To overcome this problem two models, one for the northern and one for the southern half of the corpus area were trained separately. These two different models were expected to perform better than a single model.

A.4.2 RF - Results and Feature Selection

The model for the northern half yields a mean absolute error (MAE) of 1.6232 (116.00km) with a correlation of 0.6452; the model for the southern half yields an MAE of 1.5832 (113.14km) with a correlation of 0.7383.

Using only features that had at least 1% of the maximum VI, resulted in 216 features for the northern half (with 184 Δ and $\Delta\Delta$ features) and 243 features for the southern half (with 177 Δ and $\Delta\Delta$ features).

A.4.3 SVR - Results

Using the same parameters as reported in 3.10, the SVR was able to predict the speaker locations with an MAE of 1.1932° (85.27 km) and a correlation of 0.7899 for the northern half. For the southern part of the corpus area the MAE was 1.1544° (82.50 km) with a correlation of 0.8380. This is an improvement of around 10% compared to the prediction made of the east-west direction using a single model.

A.4.4 Decision Tree

When the east-west dimension is predicted separately for the north and south half of the corpus area, the correlation rises to $R = 0.4950$ and $R = 0.5333$ and the MAE decreases to 123.02 km (1.7214°) and 127.22 km (1.7801°) in the northern and southern half of the corpus area, respectively.

Fig. A.25 shows the decision tree that was trained to predict the longitude in the east-west direction for the northern speakers of the corpus area. As the tree was trained on the complete corpus area, it also contains many Δ and $\Delta\Delta$ features high up in its hierarchy.

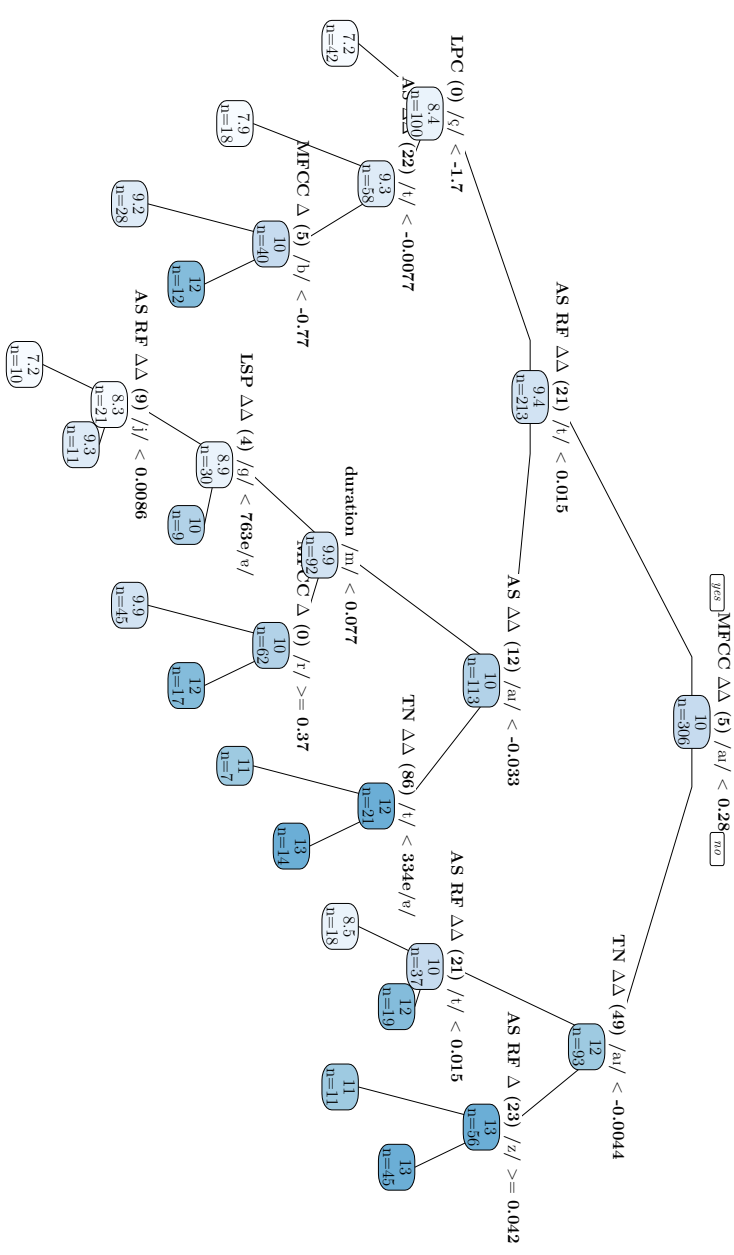


Figure A.25: Decision tree for the longitudinal direction of the northern half of the corpus area.

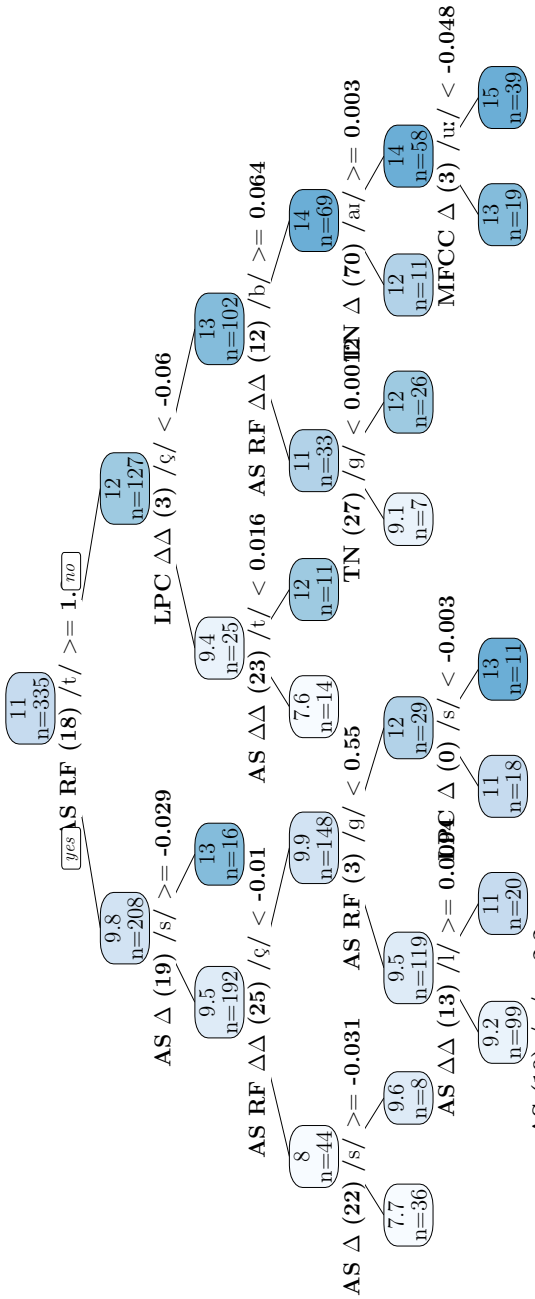


Figure A.26: Decision tree for the longitudinal direction of the southern half of the corpus area.

A.5 openSMILE Configuration

```
[componentInstances:cComponentManager]
instance[mydm].type=cDataMemory
instance[waveIn].type=cWaveSource
instance[fr1].type=cFramer
instance[w1].type=cWindower
instance[fft1].type=cTransformFFT
instance[fftmp1].type=cFFTmagphase
instance[spectral].type=cSpectral
instance[spectraldelta1].type=cDeltaRegression
instance[spectraldelta2].type=cDeltaRegression
instance[mspec].type=cMelspec
instance[audspec].type=cPlp
instance[audspecRasta].type=cPlp
instance[audspecdelta1].type=cDeltaRegression
instance[audspecdelta2].type=cDeltaRegression
instance[audspecRastadelta1].type=cDeltaRegression
instance[audspecRastadelta2].type=cDeltaRegression
instance[scale].type=cSpecScale
instance[shs].type=cPitchShs
instance[shsdelta1].type=cDeltaRegression
instance[shsdelta2].type=cDeltaRegression
instance[pitchSmooth].type=cPitchSmootherViterbi
instance[pitchSmoothdelta1].type=cDeltaRegression
instance[pitchSmoothdelta2].type=cDeltaRegression
instance[pitchSmooth2].type=cPitchSmoother
instance[harmonics].type = cHarmonics
instance[harmonicsdelta1].type=cDeltaRegression
instance[harmonicsdelta2].type=cDeltaRegression
instance[energy].type=cEnergy
instance[energydelta1].type=cDeltaRegression
instance[energydelta2].type=cDeltaRegression
instance[mzcr].type=cMZcr
instance[mzcrdelta1].type=cDeltaRegression
instance[mzcrdelta2].type=cDeltaRegression
```

```
instance[acf].type=cAcf
instance[cepstrum].type=cAcf
instance[pitchACF].type=cPitchACF
instance[pitchACFdelta1].type=cDeltaRegression
instance[pitchACFdelta2].type=cDeltaRegression
instance[mfcc].type=cMfcc
instance[mfccdelta1].type=cDeltaRegression
instance[mfccdelta2].type=cDeltaRegression
instance[tone].type=cTonespec
instance[tonedelta1].type=cDeltaRegression
instance[tonedelta2].type=cDeltaRegression
instance[chroma].type=cChroma
instance[chromadelta1].type=cDeltaRegression
instance[chromadelta2].type=cDeltaRegression
instance[lpc].type=cLpc
instance[lpcdelta1].type=cDeltaRegression
instance[lpcdelta2].type=cDeltaRegression
instance[formantLpc].type=cFormantLpc
instance[formantLpcdelta1].type=cDeltaRegression
instance[formantLpcdelta2].type=cDeltaRegression
instance[lsp].type=cLsp
instance[lspdelta1].type=cDeltaRegression
instance[lspdelta2].type=cDeltaRegression
instance[intensity].type=cIntensity
instance[intensitydelta1].type=cDeltaRegression
instance[intensitydelta2].type=cDeltaRegression
instance[pitchJitter].type=cPitchJitter
instance[pitchJitterdelta1].type=cDeltaRegression
instance[pitchJitterdelta2].type=cDeltaRegression
instance[csvSink].type = cCsvSink

[waveIn:cWaveSource]
writer.dmInstance=mydm
writer.dmLevel=wave
;filename=test/greatCut.wav
;old: filename = \cm[filename(F){test/wind.wav}:name of input file]
```

```
filename = \cm[inputfile(I):file name of the input wave file]
;filename=test/wind.wav
buffersize=160000
monoMixdown=1
sampleRate=16000

[fr1:cFramer]
reader.dmInstance=mydm
reader.dmLevel=wave
writer.dmInstance=mydm
writer.dmLevel=output
frameSize = 0.025
frameStep = 0.010
frameCenterSpecial=mid

[w1:cWindower]
reader.dmInstance=mydm
reader.dmLevel=output
writer.dmInstance=mydm
writer.dmLevel=winoutput
winFunc = ham
gain = 1.0

[fft1:cTransformFFT]
reader.dmInstance=mydm
reader.dmLevel=winoutput
writer.dmInstance=mydm
writer.dmLevel=fftc1

[fftmp1:cFFTmagphase]
reader.dmInstance=mydm
reader.dmLevel=fftc1
writer.dmInstance=mydm
writer.dmLevel=fft1

[spectral:cSpectral]
```

```
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.dmInstance=mydm
writer.dmLevel=spectral
bands[0]=250-649
bands[1]=650-999
bands[2]=1000-3999
bands[3]=4000-8000
rollOff[0] = 0.25
rollOff[1] = 0.50
rollOff[2] = 0.75
rollOff[3] = 0.90
flux=1
centroid=1
maxPos=1
minPos=1
entropy=1
variance=1
skewness=1
kurtosis=1
slope=1
harmonicity=1
sharpness=1


[spectraldelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=spectral
writer.dmInstance=mydm
writer.dmLevel=spectraldelta1

[spectraldelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=spectraldelta1
writer.dmInstance=mydm
writer.dmLevel=spectraldelta2
```

```
[mspec:cMelspec]
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.dmInstance=mydm
writer.dmLevel=mspec1
htkcompatible = 0
lofreq = 0
hifreq = 8000
showFbank = 1

; perform auditory weighting of spectrum
[audspec:cPlp]
reader.dmInstance=mydm
reader.dmLevel=mspec1
writer.dmInstance=mydm
writer.dmLevel=audspec
firstCC = 0
lpOrder = 5
cepLifter = 0
compression = 0.33
htkcompatible = 0
doIDFT = 0
doLpToCeps = 0
doLP = 0
doInvLog = 0
doAud = 1
doLog = 0
newRASTA=0
RASTA=0

; perform RASTA style filtering of auditory spectra
[audspecRasta:cPlp]
reader.dmInstance=mydm
reader.dmLevel=mspec1
writer.dmInstance=mydm
```

```
writer.dmLevel=audspecRasta
nameAppend = Rfilt
firstCC = 0
lpOrder = 5
cepLifter = 0
compression = 0.33
htkcompatible = 0
doIDFT = 0
doLpToCeps = 0
doLP = 0
doInvLog = 0
doAud = 1
doLog = 0
newRASTA=1
RASTA=0

[audspecdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=audspec
writer.dmInstance=mydm
writer.dmLevel=audspecdelta1

[audspecdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=audspecdelta1
writer.dmInstance=mydm
writer.dmLevel=audspecdelta2

[audspecRastadelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=audspecRasta
writer.dmInstance=mydm
writer.dmLevel=audspecRastadelta1

[audspecRastadelta2:cDeltaRegression]
reader.dmInstance=mydm
```

```
reader.dmLevel=audspecRastadelta1
writer.dmInstance=mydm
writer.dmLevel=audspecRastadelta2

[scale:cSpecScale]
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.levelconf.nT = 3
writer.dmInstance=mydm
writer.dmLevel=hps
// nameAppend =
copyInputName = 1
processArrayFields = 0
scale=octave
sourceScale = lin
// firstNote = 55
interpMethod = spline
minF = 25
maxF = -1
nPointsTarget = 0
specSmooth = 1
specEnhance = 1
auditoryWeighting = 1

[shs:cPitchShs]
reader.dmInstance=mydm
reader.dmLevel=hps
writer.dmInstance=mydm
writer.dmLevel=pitchShs
// nameAppend =
copyInputName = 1
processArrayFields = 0
maxPitch = 620
minPitch = 52
nCandidates = 4
scores = 1
```

```
voicing = 1
F0C1 = 0
voicingC1 = 0
F0raw = 1
voicingClip = 0
voicingCutoff = 0.700000
greedyPeakAlgo = 1
inputFieldSearch = Mag_octScale
octaveCorrection = 0
nHarmonics = 15
compressionFactor = 0.850000
lfCut = 0

[shsdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchShs
writer.dmInstance=mydm
writer.dmLevel=pitchShsdelta1

[shsdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchShsdelta1
writer.dmInstance=mydm
writer.dmLevel=pitchShsdelta2


[pitchSmooth:cPitchSmootherViterbi]
reader.dmInstance=mydm
reader2.dmInstance=mydm
reader.dmLevel=pitchShs
reader2.dmLevel=pitchShs
writer.dmInstance=mydm
writer.dmLevel=pitchG60
copyInputName = 1
bufferLength=90
F0final = 1
```

```
F0finalLog = 1
F0finalEnv = 0
voicingFinalClipped = 0
voicingFinalUnclipped = 1
F0raw = 0
voicingC1 = 0
voicingClip = 0
wTvv =10.0
wTvvd= 5.0
wTvuv=10.0
wThr = 4.0
wTuu = 0.0
wLocal=2.0
wRange=1.0

[pitchSmoothdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchG60
writer.dmInstance=mydm
writer.dmLevel=pitchG60delta1

[pitchSmoothdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchG60delta1
writer.dmInstance=mydm
writer.dmLevel=pitchG60delta2

;;;;;;;;;;;; NEW HNR
[harmonics:cHarmonics]
reader.dmInstance=mydm
reader.dmLevel = fft1;pitchG60
writer.dmInstance=mydm
writer.dmLevel = harmonics
writer.levelconf.growDyn = 0
writer.levelconf.isRb = 1
; This must be > than buffersize of viterbi smoother
```

```
writer.levelconf.nT = 200
nHarmonics = 10
f0ElementName = F0final
magSpecFieldName = pcm_fftMag
computeAcfHnrLogdB = 1

[harmonicsdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=harmonics
writer.dmInstance=mydm
writer.dmLevel=harmonicsdelta1

[harmonicsdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=harmonicsdelta1
writer.dmInstance=mydm
writer.dmLevel=harmonicsdelta2



[energy:cEnergy]
reader.dmInstance=mydm
reader.dmLevel=winoutput
writer.dmInstance=mydm
writer.dmLevel=energy

[energydelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=energy
writer.dmInstance=mydm
writer.dmLevel=energydelta1

[energydelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=energydelta1
writer.dmInstance=mydm
writer.dmLevel=energydelta2
```

```
[mzcr:cMZcr]
reader.dmInstance=mydm
reader.dmLevel=output
writer.dmInstance=mydm
writer.dmLevel=mzcr
zcr = 1
mcr = 1
amax = 0
maxmin = 0

[mzcrdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=mzcr
writer.dmInstance=mydm
writer.dmLevel=mzcrdelta1

[mzcrdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=mzcrdelta1
writer.dmInstance=mydm
writer.dmLevel=mzcrdelta2

[acf:cAcf]
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.dmInstance=mydm
writer.dmLevel=acf

[cepstrum:cAcf]
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.dmInstance=mydm
writer.dmLevel=cepstrum

[pitchACF:cPitchACF]
```

```
reader.dmInstance=mydm
reader.dmLevel=acf;cepstrum
writer.dmInstance=mydm
writer.dmLevel=pitchACF
processArrayFields = 0
voiceProb = 1
HNR = 1
F0 = 1
maxPitch = 500

[pitchACFdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchACF
writer.dmInstance=mydm
writer.dmLevel=pitchACFdelta1

[pitchACFdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=pitchACFdelta1
writer.dmInstance=mydm
writer.dmLevel=pitchACFdelta2

[mfcc:cMfcc]
reader.dmInstance=mydm
reader.dmLevel=mspec1
writer.dmInstance=mydm
writer.dmLevel=mfcc1
firstMfcc = 0
lastMfcc =  12
;the following is super important, otherwise the 0th coefficient is at the
    end (without the names changed accordingly)
htkcompatible = 0

[mfccdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=mfcc1
```

```
writer.dmInstance=mydm
writer.dmLevel=mfccdelta1

[mfccdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=mfccdelta1
writer.dmInstance=mydm
writer.dmLevel=mfccdelta2

[tone:cTonespec]
reader.dmInstance=mydm
reader.dmLevel=fft1
writer.dmInstance=mydm
writer.dmLevel=tonespec
nOctaves = 8.0

[tonedelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=tonespec
writer.dmInstance=mydm
writer.dmLevel=tonespecdelta1

[tonedelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=tonespecdelta1
writer.dmInstance=mydm
writer.dmLevel=tonespecdelta2

[chroma:cChroma]
reader.dmInstance=mydm
reader.dmLevel=tonespec
writer.dmInstance=mydm
writer.dmLevel=chroma

[chromadelta1:cDeltaRegression]
reader.dmInstance=mydm
```

```
reader.dmLevel=chroma
writer.dmInstance=mydm
writer.dmLevel=chromadelta1

[chromadelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=chromadelta1
writer.dmInstance=mydm
writer.dmLevel=chromadelta2


[lpc:cLpc]
reader.dmInstance=mydm
reader.dmLevel=output
writer.dmInstance=mydm
writer.dmLevel=lpc

[lpcdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=lpc
writer.dmInstance=mydm
writer.dmLevel=lpcdelta1

[lpcdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=lpcdelta1
writer.dmInstance=mydm
writer.dmLevel=lpcdelta2


[formantLpc:cFormantLpc]
reader.dmInstance=mydm
reader.dmLevel=lpc
writer.dmInstance=mydm
writer.dmLevel=formantLpc
```

```
[formantLpcdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=formantLpc
writer.dmInstance=mydm
writer.dmLevel=formantLpcdelta1

[formantLpcdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=formantLpcdelta1
writer.dmInstance=mydm
writer.dmLevel=formantLpcdelta2

[lsp:cLsp]
reader.dmInstance=mydm
reader.dmLevel=lpc
writer.dmInstance=mydm
writer.dmLevel=lsp

[lspdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=lsp
writer.dmInstance=mydm
writer.dmLevel=lspdelta1

[lspdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=lspdelta1
writer.dmInstance=mydm
writer.dmLevel=lspdelta2


[intensity:cIntensity]
reader.dmInstance=mydm
reader.dmLevel=output
writer.dmInstance=mydm
writer.dmLevel=intensity
```

```
[intensitydelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=intensity
writer.dmInstance=mydm
writer.dmLevel=intensitydelta1

[intensitydelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=intensitydelta1
writer.dmInstance=mydm
writer.dmLevel=intensitydelta2

;;;;;;;;;;;; taken from config file IS11_speaker_state.conf, needed for the
;;;;;;;;;;;; following jittershimmer
[pitchSmooth2:cPitchSmoother]
reader.dmInstance=mydm
reader.dmLevel=pitchShs
writer.dmInstance=mydm
writer.dmLevel=pitchF
F0raw = 0
F0final = 1
F0finalEnv = 1
voicingFinalUnclipped = 1
medianFilter0 = 0
postSmoothingMethod = simple
octaveCorrection = 0

[pitchJitter:cPitchJitter]
reader.dmInstance=mydm
reader.dmLevel = wave
writer.dmInstance=mydm
writer.dmLevel = jitterShimmer
// nameAppend =
copyInputName = 1
F0reader.dmInstance=mydm
```

```
F0reader.dmLevel = pitchF
F0field = F0final
searchRangeRel = 0.250
jitterLocal = 1
jitterDDP = 1
jitterLocalEnv = 1
jitterDDPEnv = 0
shimmerLocal = 1
shimmerLocalEnv = 0
onlyVoiced = 0

[pitchJitterdelta1:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=jitterShimmer
writer.dmInstance=mydm
writer.dmLevel=jitterShimmerdelta1

[pitchJitterdelta2:cDeltaRegression]
reader.dmInstance=mydm
reader.dmLevel=jitterShimmerdelta1
writer.dmInstance=mydm
writer.dmLevel=jitterShimmerdelta2



;;;; default (template) configuration section for component 'cCsvSink' ;;;;
[csvSink:cCsvSink]
reader.dmInstance=mydm
reader.dmLevel=energy;energydelta1;energydelta2;mzcr;mzcrdelta1;mzcrdelta2;
    pitchACF;pitchACFdelta1;pitchACFdelta2;mfcc1;mfccdelta1;mfccdelta2;chroma
    ;chromadelta1;chromadelta2;pitchShs;pitchShsdelta1;pitchShsdelta2;
    pitchG60;pitchG60delta1;pitchG60delta2;harmonics;harmonicsdelta1;
    harmonicsdelta2;jitterShimmer;jitterShimmerdelta1;jitterShimmerdelta2;
    audspec;audspecdelta1;audspecdelta2;audspecRasta;audspecRastadelta1;
    audspecRastadelta2;spectral;spectraldelta1;spectraldelta2;lpc;lpcdelta1;
    lpcdelta2;formantLpc;formantLpcdelta1;formantLpcdelta2;lsp;lspdelta1;
    lspdelta2;intensity;intensitydelta1;intensitydelta2;tonespec;
```

```
    tonespecdelta1;tonespecdelta2
;errorOnNoOutput = 0
filename = \cm[outputfile(O):file name of the output CSV file]
delimChar = ;
```

Appendix B

Second Appendix

B.1 MOCCA - Influence of Overlap Classes in the Evaluation of Automatic S&L - Experiment 2d

This experiment was conducted as an extension of the experiments in Sec. 4.5. It did not change the contents of the chapter. However, it seemed an interesting additional experiment.

When predicting the OvR, the regression algorithm has to correctly predict the OvR for feature values that represent different ways in which two segments overlap (but possess the same OvR value). The fact that two segments can overlap in different ways presumably makes this task more complicated. The different ways segments can overlap are called "overlap classes" in the following. Four different kinds of overlap classes are possible. These four different types of overlaps are shown in Figure B.1.

The reason this is likely to make the task more difficult for the regression algorithm, is the assumption that the dynamic shape of the feature values in different overlap classes is different. To test how big the influence on the result is, a simple "cheating experiment" was conducted, in which the correct overlap type is fed into the learning algorithm. The term "cheating experiment" is used, as the overlap class is not available in a real world application and it would be necessary to estimate this class before the classification (and the accuracy is unlikely to be 100%).

Experiment 2d was conducted only for the Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. The best parametrization was, analogous to Experiment 2a, found to be $C = 1$ and $\gamma = 0.1$.

Adding the correct overlap classes 1-4 as a feature leads to a small increase in the correlation coefficient. It increased from 73.36% to 74.64% (similar precision and recall). As stated before, this is likely to be less in a real world application, as this class would need to be predicted from the data and would not achieve 100% classification accuracy.

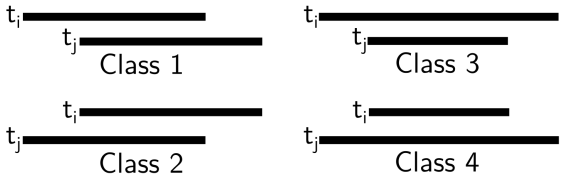


Figure B.1: The four different overlap classes for two time segments t_i and t_j.

Appendix C

Third Appendix

C.1 Previous Publications – Thesis Relevant

Thomas Kisler and Felicitas Kleber (2019). "Zur Validität automatisch segmentierter Daten. Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus". In: *Germanistische Linguistik (Marburg)*. Ed. by Sebastian Kürschner, Peter O. Müller, and Mechthild Habermann

Thomas Kisler and Florian Schiel (2018a). "MOCCA: Measure of Confidence for Corpus Analysis - Automatic Reliability Check of Transcript and Automatic Segmentation". In: *Proc. LREC*. ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA)

Thomas Kisler and Florian Schiel (2018b). "Towards a Speaker Localization from Spontaneous Speech: North-South Classification for Speakers of Contemporary German". In: *Elektronische Sprachsignalverarbeitung (ESSV) 2018 - Tagungsband der 29. Konferenz*. Vol. 29. Ulm: TUDpress, pp. 200–207. ISBN: 978-3-95908-128-3

C.2 Previous Publications – Speech Related

Thomas Kisler, Florian Schiel, and Han Sloetjes (2012). "Signal processing via web services: the use case WebMAUS". in: *Proc. DH*. Hamburg, pp. 30–34

Thomas Kisler and Uwe D. Reichel (2013a). "A dialect distance metric based on string and temporal alignment". In: *Elektronische Sprachsignalverarbeitungl*. Ed. by P. Wagner. Vol. 65. Studientexte zur Sprachkommunikation. Bielefeld: TUDpress, pp. 158–165

Thomas Kisler and Uwe D. Reichel (2013b). "Exploring the connection of acoustic and distinctive features". In: *Proc. Interspeech*. Lyon, pp. 320–324

U. D. Reichel and T. Kisler (2014). "Language-independent grapheme-phoneme conversion and word stress assignment as a web service". In: *Elektronische Sprachverarbeitung 2014*. Ed. by R. Hoffmann. Vol. 71. Studientexte zur Sprachkommunikation. Dresden, Germany: TUDpress, pp. 42–49

Florian Schiel and Thomas Kisler (2014). "German Alcohol Language Corpus - the Question of Dialect". In: *Proc. LREC*. ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 353–356. ISBN: 978-2-9517408-8-4

Thomas Kisler, Florian Schiel, Uwe D. Reichel, and Christoph Draxler (2015). "Phonetic/linguistic web services at BAS". in: *Proc. Interspeech*. Dresden, Germany, paper 2609

Thomas Kisler, Uwe Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl, and Nina Pörner (2016). "BAS Speech Science Web Services - an Update of Current Developments". In: *Proc. LREC*. ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1

Uwe D. Reichel, Florian Schiel, Thomas Kisler, and Nina Pörner (2016). "The BAS Speech Data Repository". In: *Proc. LREC*. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1. URL: https://www.phonetik.uni-muenchen.de/forschung/publikationen/ReichelSchielKislerDraxlerPoerner-LREC2016.pdf

Thomas Kisler, Uwe Reichel, and Florian Schiel (2017). "Multilingual processing of speech via web services". In: *Computer Speech & Language*. ISSN: 0885-2308. DOI: http://dx.doi.org/10.1016/j.csl.2017.01.005

Zusammenfassung

Zusammenfassung auf Deutsch

Nachfolgend werden die drei großen Experimente, die in den Hauptkapiteln dieser Dissertation beschrieben werden, in deutscher Sprache zusammengefasst.

Zur Validität automatisch segmentierter Daten

In dieser Untersuchung wurde das dialektale Merkmal der komplementären Länge im Bairischen herangezogen, um eine Aussage über die Validität der automatisch segmentierten Daten zu treffen. Dieses dialektale Merkmal lässt sich durch das Verhältnis der Länge des Vokals zur Länge des Vokals plus des Konsonanten, dem V/(V+K)-Verhältnis (Kohler, 1979), beschreiben.

Die automatische S&E, die überprüft werden sollte, wurde mithilfe von WebMAUS (Kisler et al., 2017) erzeugt. Die Analyse basiert auf einem Teil der Maptask-Daten (vgl. Anderson et al., 1991) des Deutsch Heute Korpus (Brinckmann et al., 2008). Der verwendete Teil umfasst insgesamt 87 Sprecher des Korpus, davon 42 ost- und 22 westmittelbairische Sprecher (unterteilt nach Wiesinger (1990)) und 23 ostfränkische Sprecher.

Es wurde gezeigt, dass mit den automatisch segmentierten Daten eine valide Analyse möglich ist. Die Verteilung der Ausprägungen des V/(V+K)-Verhältnisses zeigt, wie erwartet, eine komplementäre Verteilung der Länge in den mittelbairischen Sprechern und eine relative freie Kombinierbarkeit der Lang-/Kurzvokale und der Lenis-/Fortiskonsonanten in den OF Sprechern. Zudem deuten die Verteilungen des V/(V+K)-Verhältnisses in den WMB Sprechern tatsächlich auf einen vermeintlichen Wandel hin, da dort auch Langvokale

vor Fortisplosiven auftauchen können (wie z.B. berichtet in Kleber, 2017). Die Ausdehnung der Verteilung des V/(V+K)-Verhältnisses über einen großen Teil des Wertebereichs lässt auf ein gewisses Rauschen bei der automatischen S&E schließen.

In einem zweiten Experiment wurde ein Subset von 56 Sprechern aller Gebiete manuell nachkorrigiert, um das vom automatischen Alignment eingeführte Rauschen zu überprüfen. Basierend auf den Segmentgrenzen des korrigierten Subkorpus urde gezeigt, dass a) die Ausdehnung der Verteilung des V/(V+K)-Verhältnisses über den Wertebereich kleiner wird und b) sich die verschiedenen Kategorien innerhalb einer Gruppe besser gegeneinander abgrenzen lassen. Die Ergebnisse des unkorrigierten Datensatzes korrelieren mit $R = 0.58$ mit den Ergebnissen basierend auf dem korrigierten Datensatz. Dies bedeutet, dass die ursprüngliche automatische S&E somit zur Untersuchung von Dialektmerkmalen geeignet ist, durch eine manuelle Nachkorrektur aber noch feinere Unterschiede herausgestellt werden können.

Geolokalisierung der Sprecherherkunft

In der Dialektologie des Deutschen geht man von einer semi-kontinuierlichen Veränderung der Dialektmerkmale aus (vgl. z.B. Haag, 1929, S. 19, Barbour et al., 1990, S. 136). Dieses Experiment hatte zum Ziel diese Veränderung über das gesamte deutschsprachige Gebiet anhand von akustischen Features und vorhersagebasierten Modellen zu analysieren.

Für die Modellierung wurden die gesamten Maptask-Aufnahmen des Deutsch Heute Korpus verwendet, die automatisch von WebMAUS verarbeitet werden konnten. Dies entspricht dem Material von 641 Gewährspersonen (313 männlich, 328 weiblich) und ungefähr 67 h Sprache, wobei die folgenden 41 Phoneme im Korpus nach automatischer S&E auftreten: /ə, a, ɐ, aː, aɪ, aʊ, b, ç, d, e, eː, ɛ, ɛː, f, g, h, i, iː, ɪ, j, k, l, m, ŋ, o, oː, ɔ, p, r, s, ʃ, t, u, uː, ʊ, v, w, x, yː, ʏ, z/.

Von diesem Sprachmaterial wurden mithilfe von openSMILE (Eyben et al., 2010) insgesamt 731 Features extrahiert, wobei das Featureset auch Δ- und ΔΔ-Features enthält. Die Features wurden zum Phonemmittelpunkt extrahiert und über die 20% Region (Mittelpunkt ±10%) gemittelt. Auf diesem Datensatz wurden drei Experimente durchgeführt,

wobei allen Experimenten ein Cross Validation (CV) Testverfahren zugrundeliegt, das sicherstellt, dass jeder Sprecher, der zum Test dient, dem System bis dahin unbekannt war.

Das erste Experiment basiert auf Random Forests (RFs), die auf Basis der extrahierten Features eine Vorhersage machen aus welcher Hälfte Deutschlands ein bestimmter Sprecher stammt. Für jedes vorhandene Phonem und beide Richtungen (Ost/West und Nord/Süd) wird ein Vorhersagemodell trainiert. Für die Ost/West Erkennung ist die Vorhersage mit einer Accuracy von 0.5791 auf dem Phonem /øː/ möglich (für 31 Phoneme ist eine Vorhersage über dem Zufall möglich) und für Nord/Süd eine Erkennung mit einer Accuracy von 0.7037 auf Phonem /z/ (für alle Phoneme ist eine Vorhersage über dem Zufall möglich). Es wurden für beide Hälften die Phoneme, für die die besten Resultate erzielt werden konnten, mit bereits bekannten Dialektmerkmalen in Verbindung gebracht.

Das zweite Experiment ist eine Wiederholung des ersten Experiments mit dem Unterschied, dass keine kategoriale Variable vorhergesagt wird sondern ein kontinuierlicher Wert. Die Vorhersage erfolgte ebenfalls wieder getrennt für beide Richtungen – Ost/West und Nord/Süd. Da eine solche kontinuierliche Vorhersage noch nie durchgeführt wurde, musste ein Vergleichswert gefunden werden, um ein nützliches[1] Modell von einem nutzlosen zu unterscheiden. Dafür wurde ein Nullmodel (James et al., 2014, S. 205) definiert, das für jeden Sprecher den Mittelpunkt der Aufnahmeorte des verwendeten Deutsch Heute Korpus als Vorhersage zurückgibt. Durch die Vorhersage mit RFs ist für 40 Phoneme in Ost/West und 31 Phonemen in Nord/Süd eine Vorhersage möglich, die besser ist als der Vergleichswert. Jedoch erreicht auch das beste Phonem /z/ in Nord/Süd-Richtung nur eine verbesserte Vorhersage von 26.69 km gegenüber dem konservativ gewählten Vergleichswert. Basierend auf den guten Ergebnissen der Klassifikation ist das überraschend. Für die Ost/West-Richtung ist es nur eine Verbesserung um 9.45 km.

Im dritten Experiment wurde untersucht, wie sich die Aggregation von mehreren Realisierungen eines Phonems pro Sprecher und darüber hinaus mehrerer Phoneme auf die Erkennungsleistung auswirkt und wie sich der geografische Raum damit aufteilen lässt. Für dieses Experiment wurde ein sehr großes Featureset erzeugt, in dem zuerst über alle Phoneme eines Sprechers gemittelt wird und anschließend die Features aller 33 Phoneme,

[1] Hier wird *Nutzen* mit einer Verbesserung der Vorhersageleistung in jeglicher Form definiert.

die von allen Sprechern geäußert wurden, konkateniert werden. Durch eine anschließende Featureselektion und die Verwendung eines Support Vector Regression (SVR) Modells konnte die Lokalisierung erheblich gesteigert werden. In Ost/West-Richtung um 55.3 km in Nord/Süd-Richtung um 113.95 km gegenüber dem Nullmodel. Zusätzlich wurde in diesem Experiment gezeigt, dass sich die Interpretation der Features im geografischen Raum basierend auf einer Vorhersage mit RFs, besonders in Nord-Süd-Richtung, lohnen kann. Das ist der Fall, da sich die meisten Aufteilungen, die auf verschiedenen akustischen Features von verschiedenen Phonemen basieren, zumindest theoretisch – also ohne manuelle ohrenphonetische Validierung aller realisierten Phoneme aller Sprecher – mit dialektalem Wissen in Verbindung bringen lassen.

Measure of Confidence for Corpus Analysis (MOCCA)

In der Erstellung von Korpora für die phonetische Forschung gibt es einige Teilschritte die oft mit Fehlern oder großen Abweichungen verbunden sind. Im Einzelnen sind das die orthografische Transkription und die automatische S&E. Es wurde ein System entwickelt, dass diese Schritte unterstützt, das auf Verfahren basiert, die bei der Qualitätsbestimmung von automatischer Spracherkennung bereits ihren Nutzen bewiesen haben.

Als Datenbasis dienten die semi-spontansprachlichen Aufnahmen des Kiel-Corpus (Kohler, 1996; John, 2012) von 30 Sprechern (2225 Äußerungen) und die gelesene Sprache des PhonDat 2 Corpus (The ASR Consortium, 1995) von 16 Sprechern (1024 Äußerungen). Dabei diente der Kiel-Corpus als Trainingskorpus zur Hyperparametersuche und der PhonDat 2 Korpus als unabhängiges Testkorpus, wobei beide eine manuelle S&E besitzen.

Die Vorhersage basiert auf einem Subset der Features aus Schaaf et al. (1997) und waren diejenigen, die im Rahmen des MAUS Systems ohne zusätzlichen Modellierungsaufwand zur Verfügung standen. Da in MAUS die Modellierung auf Phonebene stattfindet, sorgt das bei einer Vorhersage auf Wortebene zu variablen Featurevektorlängen. Um dieses Problem zu umgehen, wurden Maße der Basisfeatures herangezogen, die die Lage und Dynamik der Werte beschreiben. Das waren im Einzelnen die Summe, der Mittelwert, der Median, die Spannweite, die Varianz, die Standardabweichung und die ersten drei Koeffizienten einer

diskreten Kosinustransformation. Dadurch konnte eine konstante Featurevektorlänge von $d = 30$ erreicht werden. Diese Featurevektoren wurden zusammen mit RFs und SVR für die Vorhersage benutzt.

Das erste Experiment konzentrierte sich auf die Auffindung von falsch transkribierten Wörtern (oder Wörtern, die nicht zum Signal passen, z.B. aufgrund von einer orthografischen Verschriftung von dialektal stark gefärbtem Material). Da im Korpus keine Verschriftungsfehler bekannt sind, wurde eine Ersetzungsstrategie angewandt, um so künstliche Fehler an bekannten Positionen zu erzeugen, damit das Training durchgeführt werden konnte. Es wurde gezeigt, dass es mit einem SVR-Modell möglich ist falsch transkribierte Wörter im unabhängigen Testset mit einer Accuracy von 0.7876 zu bestimmen. Die Ausgabe des Systems ist dabei ein kontinuierlicher Wert, der aussagt wie hoch der Klassifikator die Klassenzugehörigkeitswahrscheinlichkeit des momentanen Wortes einschätzt. Mithilfe eines Schwellwerts wurde dieser Wert auf einen kategorialen Wert abgebildet, der dann schließlich signalisiert, ob es sich um ein korrekt oder falsch transkribiertes Wort handelt.

Das zweite Experiment hatte zum Ziel, diejenigen Stellen zu identifizieren, die durch das automatische Alignment nicht oder nur unzureichend genau festgestellt werden konnten. Dazu wurde für jedes Wort im Trainingskorpus das Überlappungsverhältnis zwischen automatisch generierter und manuell erzeugter phonetischer Transkription berechnet. Dieses Überlappungsverhältnis war der Wert, den das System vorhersagen sollte. Da die Anzahl der Messungen über den Wertebereich sehr unterschiedlich war (heteroskedastischer Fehler), wurde eine kombinierte Unter- und Übersamplingstrategie an den entsprechenden Stellen im Wertebereich des Überlappungsverhältnisses angewandt. Durch diese Strategie konnte ein gleichmäßigerer (homoskedastischerer) Fehler über den Wertebereich erzielt werden. Mit dem SVR-Modell war eine Vorhersage auf dem unabhängien Testset möglich, die mit den tatsächlichen Werten mit $R = 0.60$ korreliert.

Insgesamt liegen die Erkennungsraten für falsch transkribierte Wörter und stark abweichende automatische S&E in einem Bereich, in dem sie einen sinnvollen Beitrag zur Auffindung von Fehlern in großen Korpora leisten können. Hier ist davon aus zugehen, dass die Größe von Korpora in der Zukunft eher noch steigen wird und damit die Notwendigkeit von automatischer Fehlererkennung.

Bibliography

Anderson, Anne H, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. (1991). "The HCRC map task corpus". In: *Language and Speech* 34.4, pp. 351–366.

Archer, Kellie J and Ryan V Kimes (2008). "Empirical characterization of random forest variable importance measures". In: *Computational Statistics & Data Analysis* 52.4, pp. 2249–2260.

Arslan, Levent M and John HL Hansen (1996). "Language accent classification in American English". In: *Speech Communication* 18.4, pp. 353–367.

Auer, Peter and Frans Hinskens (1996). "The convergence and divergence of dialects in Europe. New and not so new developments in an old area". In: *Sociolinguistica* 10, pp. 1–30.

Bahari, Mohamad Hasan, Rahim Saeidi, Hugo Van hamme, and David Van Leeuwen (2013). "Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech". In: *Proc. ICASSP*. IEEE, pp. 7344–7348. DOI: 10.1109/ICASSP.2013.6639089.

Bannert, Robert (1976). *Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage*. Vol. 10. CWK Gleerup.

Barbour, Stephen and Patrick Stevenson (1990). *Variation in German: A critical approach to German sociolinguistics*. Cambridge University Press.

Bauer, Roland (2004). "Dialekte–Dialektmerkmale–dialektale Spannungen. Von 'Cliquen','Störenfrieden'un 'Sündenböcken'im Netz des dolomitenladinischen Sprachatlasses ALD-I". In: *Ladinia* 28, pp. 201–242.

Becker, Thomas (1998). *Das Vokalsystem der deutschen Standardsprache*. Ed. by Konrad Ehlich. Frankfurt am Main.

Beddor, Patrice Speeter (2009). "A coarticulatory path to sound change". In: *Language* 85.4, pp. 785–821.

Biadsy, Fadi (2011). "Automatic dialect and accent recognition and its application to speech recognition". PhD thesis.

Biadsy, Fadi, Julia Hirschberg, and Michael Collins (2010). "Dialect Recognition using Phone-GMM-Supervector-based SVM Kernel". In: *Proc. Interspeech*.

Borland, D. and R. M. Taylor Ii (2007). "Rainbow Color Map (Still) Considered Harmful". In: *IEEE Computer Graphics and Applications* 27.2, pp. 14–17. ISSN: 0272-1716. DOI: 10.1109/MCG.2007.323435.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). "A Training Algorithm for Optimal Margin Classifiers". In: *Proc. ACM Workshop on Computational Learning Theory*. ACM Press, pp. 144–152.

Braunschweiler, Norbert (1997). "Integrated cues of voicing and vowel length in German: A production study". In: *Language and Speech* 40.4, pp. 353–376.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Brinckmann, Caren, Stefan Kleiner, Ralf Knöbl, and Nina Berend (2008). "German Today: an areally extensive corpus of spoken Standard German". In: *Proc. LREC*.

Brown, Georgina (2015). "Automatic Recognition of Geographically-Proximate Accents Using Content-Controlled and Content-Mismatched Data". In: *Proc. ICPhS*. Ed. by The Scottish Consortium for ICPhS 2015. Glasgow, UK.

Burger, Susanne and Florian Schiel (1998). "RVG 1-A Database for Regional Variants of Contemporary German". In: *Proc. LREC*, pp. 1083–1087.

Campbell, William M, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo (2006). "Support vector machines for speaker and language recognition". In: *Computer Speech & Language* 20.2–3, pp. 210–229.

Chambers, Jack K. and Peter Trudgill (1998). *Dialectology*. Cambridge Textbooks in Linguistics. Cambridge University Press. ISBN: 9780521596466.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: a library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3, p. 27.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, W., S. Ananthakrishnan, R. Kumar, R. Prasad, and P. Natarajan (2013). "ASR error detection in a conversational spoken language translation system". In: *Proc. ICASSP*, pp. 7418–7422. DOI: 10.1109/ICASSP.2013.6639104.

Clauss, Günter and Heinz Ebner (1974). *Grundlagen [der Statistik]*. Volkseigner Verlag.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1, pp. 37–46.

Cox, Stephen and Richard Rose (1996). "Confidence measures for the switchboard database". In: *Proc. ICASSP*. Vol. 1. IEEE, pp. 511–514.

Crystal, Thomas H and Arthur S House (1990). "Articulation rate and the duration of syllables and stress groups in connected speech". In: *Journal of the Acoustical Society of America* 88.1, pp. 101–112.

Cunha, C., J. Harrington, S. Moosmüller, and J. Brandstätter (2015). "The influence of consonantal context on the tense-lax contrast in two standard varieties of German". In: *Trends in phonetics and phonology in German speaking Europe*. Frankfurt am Main: Peter Lang, pp. 65–77.

DeMarco, Andrea and Stephen J Cox (2013). "Native accent classification via i-vectors and speaker compensation fusion." In: *Proc. Interspeech*, pp. 1472–1476.

Díaz-Uriarte, Ramón and Sara Alvarez De Andres (2006). "Gene selection and classification of microarray data using random forest". In: *BMC bioinformatics* 7.1, p. 3.

Draxler, Christoph and Susanne Burger (1997). "Identification of regional variants of high German from digit sequences in German telephone speech". In: *Proc. Eurospeech*.

Draxler, Christoph and Florian Schiel (2016). *We talked about how much time an orthographic transcription needs compared to a phonetic transcription.* Personal Communication. München, Germany.

Drucker, Harris, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik (1997). "Support Vector Regression Machines". In: *Advances in Neural Information Processing Systems 9*. Ed. by M. C. Mozer, M. I. Jordan, and T. Petsche. MIT Press, pp. 155–161. URL: http://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf.

Duden Online (2018). https://www.duden.de/, (last accessed October 7, 2018).

D'Arcy, Shona M., Martin J. Russell, Sue R. Browning, and Mike J. Tomlinson (2004). "The accents of the British Isles (ABI) corpus". In: *Proceedings Modélisations pour l'Identification des Langues*, pp. 115–119.

Evermann, Gunnar and Philip C Woodland (2000). "Large vocabulary decoding and confidence estimation using word posterior probabilities". In: *Proc. ICASSP*. Vol. 3. IEEE, pp. 1655–1658.

Eyben, Florian, Martin Wöllmer, and Björn Schuller (2010). "openSMILE: the Munich versatile and fast open-source audio feature extractor". In: *Proc. ACMMM*. MM '10. Firenze, Italy: ACM, pp. 1459–1462. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246. URL: http://doi.acm.org/10.1145/1873951.1874246.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *Journal of Machine Learning Research* 15.1, pp. 3133–3181.

Finkelstein, Samantha, Amy Ogan, Caroline Vaughn, and Justine Cassell (2013). "Alex: A virtual peer that identifies student dialect". In: *Proc. Culturally-aware Technology Enhanced Learning in conjuction with EC-TEL*.

Ghannay, Sahar, Nathalie Camelin, and Yannick Estève (2015). "Which ASR errors are hard to detect?" In: *Proc. ERRARE*. Sinaia (Romania).

Gillick, Larry, Yoshiko Ito, and Jonathan Young (1997). "A probabilistic approach to confidence estimation and evaluation". In: *Proc. ICASSP*. Vol. 2. IEEE, pp. 879–882.

Goebl, Hans (2010). "Dialectometry and quantitative mapping". In: *An International Handbook of Linguistic Variation*. Ed. by Alfred Lameli Roland Kehrein Stefan Rabanus. Vol. Volume 2: Language Mapping. De Gruyter Mouton.

Gooskens, Charlotte and Wilbert Heeringa (2004). "The position of Frisian in the Germanic language area". In: *On the boundaries of phonology and phonetics*, pp. 61–87.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts (2013). "A multivariate spatial analysis of vowel formants in American English". In: *Journal of Linguistic Geography* 1.1, pp. 31–51.

Gussenhoven, Carlos (2004). *The Phonology of Tone and Intonation*. Cambridge University Press.

Guyon, Isabelle and André Elisseeff (Mar. 2003). "An Introduction to Variable and Feature Selection". In: *J. Mach. Learn. Res.* 3, pp. 1157–1182. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944968.

Haag, Karl (1929). "Sprachwandel im Lichte der Mundartgrenzen". In: *Teuthonista* 6.1, pp. 1–35. ISSN: 08635781. URL: http://www.jstor.org/stable/40498732.

Hanani, A., M.J. Russell, and M.J. Carey (2013). "Human and computer recognition of regional accents and ethnic groups from British English speech". In: *Computer Speech & Language* 27.1, pp. 59–74. ISSN: 0885-2308.

Hansen, John H. L., Umit H. Yapanel, Rongqing Huang, and Ayako Ikeno (2004). "Dialect analysis and modeling for automatic classification". In: *Proc. Interspeech*.

Harrington, Jonathan, Felicitas Kleber, and Ulrich Reubold (2008). "Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study". In: *Journal of the Acoustical Society of America* 123.5, pp. 2825–2835.

Hastie, Tevor, Robert Tibshirani, and Jerome Friedman (2013). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Second. Springer.

Hastie, Trevor (2014). *Trevor Hastie – Gradient Boosting Machine Learning*. https://www.youtube.com/watch?v=wPqtzj5VZus, last accessed September 06, 2018.

Hawkins, Sarah (2003). "Roles and representations of systematic fine phonetic detail in speech understanding". In: *Journal of Phonetics* 31.3-4, pp. 373–405.

Heeringa, Wilbert and Charlotte Gooskens (2003). "Norwegian Dialects Examined Perceptually and Acoustically". In: *Computers and the Humanities* 37.3, pp. 293–315. ISSN:

1572-8412. DOI: 10.1023/A:1025087115665. URL: https://doi.org/10.1023/A:1025087115665.

Heeringa, Wilbert, Keith Johnson, and Charlotte Gooskens (Feb. 2009). "Measuring Norwegian dialect distances using acoustic features". In: *Speech Communication* 51.2, pp. 167–183. ISSN: 0167-6393. DOI: 10.1016/j.specom.2008.07.006. URL: http://dx.doi.org/10.1016/j.specom.2008.07.006.

Hermansky, H. and N. Morgan (1994). "RASTA processing of speech". In: *IEEE Transactions on Speech and Audio Processing* 2.4, pp. 578–589. ISSN: 1063-6676. DOI: 10.1109/89.326616.

Hillenbrand, James and Robert T. Gayvert (1993). "Vowel classification based on fundamental frequency and formant frequencies". In: *Journal of Speech, Language, and Hearing Research* 36.4, pp. 694–700.

Hinderling, Robert, Werner König, Ludwig M. Eichinger, Hans-Werner Eroms, Horst Haider Munske, and Norbert Richard Wolf (1996 – 2014). *Bayerischer Sprachatlas*.

Hinrichs, Erhard W., Marie Hinrichs, and Thomas Zastrow (2010). "WebLicht: Web-Based LRT Services for German". In: *Proc. ACL*, pp. 25–29. URL: http://www.aclweb.org/anthology/P10-4005.

Huckvale, Mark (2004). "ACCDIST: a metric for comparing speakers' accents". In: *Proc. Interspeech*, pp. 29–32.

— (2007). "ACCDIST: an accent similarity metric for accent recognition and diagnosis". In: *Speaker Classification II*. Springer, pp. 258–275.

Irizarry, Rafael (2017). *The joy of no more violin plots*. https://simplystatistics.org/2017/07/13/the-joy-of-no-more-violin-plots/, (last accessed October 7, 2018).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2014). *An introduction to statistical learning*. Vol. 6. Springer.

Jiang, Hui (2005). "Confidence measures for speech recognition: A survey". In: *Speech Communication* 45.4, pp. 455–470.

John, Tina (2012). "EMU Speech Database System: praxisorientierte Weiterentwicklung der Funktionalität, Benutzerfreundlichkeit und Interoperabilität sowie die Aufbere-

itung des Kiel Corpus als EMU-Sprachdatenbank". PhD thesis. Ludwig-Maximilians-Universität München.

Johnson, Keith (2011). *Acoustic and Auditory Phonetics*. 3rd. Wiley-Blackwell.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education Upper Saddle River.

Kamppari, Simo O and Timothy J Hazen (2000). "Word and phone level acoustic confidence scoring". In: *Proc. ICASSP*. Vol. 3. IEEE, pp. 1799–1802.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik, and Achim Zeileis (2004). "kernlab – An S4 Package for Kernel Methods in R". In: *Journal of Statistical Software* 11.9, pp. 1–20. URL: `http://www.jstatsoft.org/v11/i09/`.

Karatzoglou, Alexandros, David Meyer, and Kurt Hornik (2006). "Support vector machines in R". In: *Journal of Statistical Software* 15.09.

Kat, Liu Wai and Pascale Fung (1999). "Fast accent identification and accented speech recognition". In: *Proc. ICASSP*. Vol. 1. IEEE, pp. 221–224.

Kemp, Thomas and Thomas Schaaf (1997). "Estimating confidence using word lattices." In: *Proc. Eurospeech*, pp. 827–830.

Kipp, Andreas, Maria-Barbara Wesenick, and Florian Schiel (1997). "Pronuncation modeling applied to automatic segmentation of spontaneous speech". In: *Proc. Eurospeech*. Ed. by George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas. Rhodes, Greece, 1023–1026.

Kisler, Thomas, Florian Schiel, and Han Sloetjes (2012). "Signal processing via web services: the use case WebMAUS". In: *Proc. DH*. Hamburg, pp. 30–34.

Kisler, Thomas and Uwe D. Reichel (2013a). "A dialect distance metric based on string and temporal alignment". In: *Elektronische Sprachsignalverarbeitungl*. Ed. by P. Wagner. Vol. 65. Studientexte zur Sprachkommunikation. Bielefeld: TUDpress, pp. 158–165.

— (2013b). "Exploring the connection of acoustic and distinctive features". In: *Proc. Interspeech*. Lyon, pp. 320–324.

Kisler, Thomas and Florian Schiel (2014). *Towards understanding the regional distribution of acoustic features from speech*. https://methodsxv.webhosting.rug.nl/abstracts/All/Posters/KISLERsCHIEL.pdf.

Kisler, Thomas, Florian Schiel, Uwe D. Reichel, and Christoph Draxler (2015). "Phonetic/linguistic web services at BAS". In: *Proc. Interspeech*. Dresden, Germany, paper 2609.

Kisler, Thomas, Uwe Reichel, Florian Schiel, Christoph Draxler, Bernhard Jackl, and Nina Pörner (2016). "BAS Speech Science Web Services - an Update of Current Developments". In: *Proc. LREC*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Kisler, Thomas, Uwe Reichel, and Florian Schiel (2017). "Multilingual processing of speech via web services". In: *Computer Speech & Language*. ISSN: 0885-2308. DOI: http://dx.doi.org/10.1016/j.csl.2017.01.005.

Kisler, Thomas and Florian Schiel (2018a). "MOCCA: Measure of Confidence for Corpus Analysis - Automatic Reliability Check of Transcript and Automatic Segmentation". In: *Proc. LREC*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA).

— (2018b). "Towards a Speaker Localization from Spontaneous Speech: North-South Classification for Speakers of Contemporary German". In: *Elektronische Sprachsignalverarbeitung (ESSV) 2018 - Tagungsband der 29. Konferenz*. Vol. 29. Ulm: TUDpress, pp. 200–207. ISBN: 978-3-95908-128-3.

Kisler, Thomas and Felicitas Kleber (2019). "Zur Validität automatisch segmentierter Daten. Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus". In: *Germanistische Linguistik (Marburg)*. Ed. by Sebastian Kürschner, Peter O. Müller, and Mechthild Habermann.

Klatt, Dennis H (1973). "Interaction between two factors that influence vowel duration". In: *Journal of the Acoustical Society of America* 54.4, pp. 1102–1104.

Kleber, Felicitas (2017). "Complementary length in vowel-consonant sequences: acoustic and perceptual evidence for a sound change in progress in Bavarian German". In: *Journal of the International Phonetics Association*, pp. 1–22. DOI: 10.1017/S0025100317000238.

— (2018). *We talked about variation that is known in Germany and might explain the features used for splitting decisoin trees trained on the Digital Humanities (DH) corpus.* Personal Communication. München, Germany.

Kleber, Felicitas, Tina John, and Jonathan Harrington (2010). "The implications for speech perception of incomplete neutralization of final devoicing in German". In: *Journal of Phonetics* 38.2, pp. 185–196.

Kleiner, Stefan, Nina Berend, Caren Brinckmann, and Ralf Knöbl (2007). "" Deutsch heute": ein sprachgebietsweites Forschungsprojekt zur regionalen Variation in der gesprochenen deutschen Standardsprache". In: ed. by Heinz-Dieter Pohl.

Kohler, Klaus et al. (1995). *Das Kiel Corpus.*

Kohler, Klaus J. (1977). "The production of plosives". In: *Arbeitsberichte des Instituts für Phonetik der Universität Kiel* 8, pp. 30–110.

— (1979). "Dimensions in the perception of fortis and lenis plosives". In: *Phonetica* 36.4-5, pp. 332–343.

— (1996). "Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech". In: *Proc. ICSLP*. Vol. 3. IEEE, pp. 1938–1941.

König, Werner (1989). *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland: Text.* Vol. 1 (Text). M. Hueber Verlag.

— (2005). *dtv-Atlas Deutsche Sprache.* 13th ed. Deutscher Taschenbuch Verlag.

Kufner, Herbert L. (1964). *München.* Göttingen: Vandenhoeck & Ruprecht.

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. (2017). *caret: Classification and Regression Training.* URL: https://CRAN.R-project.org/package=caret.

Lameli, Alfred (2008a). “Deutsche Sprachlandschaften”. In: *Nationalatlas aktuell* 9.

— (2008b). “Was Wenker noch zu sagen hatte... Die unbekannten Teile des "Sprachatlas des Deutschen Reichs"”. In: *Zeitschrift für Dialektologie und Linguistik*, pp. 255–281.

— (2013). *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Vol. 54. Walter de Gruyter.

Landesbibliothek, Online Bayerische (2013). *Sprechender Sprachatlas von Bayern*. `https://sprachatlas.bayerische-landesbibliothek-online.de/` (last accessed August 10, 2018).

Larcher, Anthony, Jean-Francois Bonastre, Benoit G.B. Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John SD Mason, and Jean-Yves Parfait (2013). “ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition”. In: *Proc. Interspeech*, pp. 2768–2772.

Leemann, Adrian, Marie-José Kolly, and David Britain (2018). “The English Dialects App: the creation of a crowdsourced dialect corpus”. In: *Ampersand* 5, pp. 1–17. ISSN: 2215-0390. DOI: `10.1016/j.amper.2017.11.001`.

Levenshtein, V. I. (Feb. 1966). “Binary Codes Capable of Correcting Deletions, Insertions and Reversals”. In: *Soviet Physics Doklady* 10, p. 707.

Liaw, Andy and Matthew Wiener (2002a). “Classification and Regression by randomForest”. In: *R News* 2.3, pp. 18–22.

— (2002b). “Classification and regression by randomForest”. In: *R news* 2.3, pp. 18–22.

Llompart, Miquel and Eva Reinisch (2017). “Articulatory information helps encode lexical contrasts in a second language.” In: *Journal of Experimental Psychology: Human Perception and Performance* 43.5, p. 1040.

Löffler, Heinrich (2003). *Dialektologie: eine Einführung*. Gunter Narr Verlag.

Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno (2014). “Automatic language identification using deep neural networks”. In: *Proc. ICASSP*. IEEE, pp. 5337–5341.

Louppe, Gilles (2014). “Understanding Random Forests: From Theory to Practice”. PhD thesis.

Machelett, Kirsten (1996). *Das Lesen von Sonagrammen V1.0 - Kapitel III - Vorlesungsunterlagen*. https://www.phonetik.uni-muenchen.de/studium/skripten/SGL/SGLKap3.html (last accessed August 10, 2018).

Malley, James D., Jochen Kruppa, Abhijit Dasgupta, Karen G Malley, and Andreas Ziegler (2012). "Probability machines: consistent probability estimation using nonparametric learning machines". In: *Methods of Information in Medicine* 51.1, p. 74.

Mangu, Lidia, Eric Brill, and Andreas Stolcke (2000). "Finding consensus in speech recognition: word error minimization and other applications of confusion networks". In: *Computer Speech & Language* 14.4, pp. 373–400.

Mathussek, Andrea (2016). "On the problem of field worker isoglosses". In: *The future of dialects: Selected papers from Methods in Dialectology XV*. Ed. by John Nerbonne, Marie-Hélène Côté, and Remco Knooihuize. Vol. Language Variation 1. Berlin: Language Science Press. Chap. 6, pp. 99–115. URL: http://langsci-press.org/catalog/book/81.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. URL: https://CRAN.R-project.org/package=e1071.

Montaño, Raúl and Francesc Alías (2017). "The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages". In: *Speech Communication* 88, pp. 1–16.

Moosmüller, Sylvia and Julia Brandstätter (2014). "Phonotactic information in the temporal organization of Standard Austrian German and the Viennese dialect". In: *Language Sciences* 46, pp. 84–95.

More, Ajinkya (2016). "Survey of resampling techniques for improving classification performance in unbalanced datasets". In: *arXiv preprint arXiv:1608.06048*.

Moreland, Kenneth (2009). "Diverging color maps for scientific visualization". In: *International Symposium on Visual Computing*. Springer, pp. 92–103.

Murty, K. S. R. and B. Yegnanarayana (2006). "Combining evidence from residual phase and MFCC features for speaker recognition". In: *IEEE Signal Processing Letters* 13.1, pp. 52–55. ISSN: 1070-9908. DOI: 10.1109/LSP.2005.860538.

Najafian, Maryam (2016). "Acoustic model selection for recognition of regional accented speech". PhD thesis. University of Birmingham.

Najafian, Maryam, Saeid Safavi, Phil Weber, and Martin Russell (2016). "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems". In: *Proc. Odyssey*.

National Imagery and Mapping Agency (2000). *Department of Defense World Geodetic System 1984 - It's Definition and Relationships with Local Geodetic Systems*. Tech. rep. Reston, US: National Imagery and Mapping Agency.

Nembrini, Stefano, Inke R. König, and Marvin N. Wright (2018). "The revival of the Gini importance?" In: *Bioinformatics*, bty373. DOI: 10.1093/bioinformatics/bty373. URL: http://dx.doi.org/10.1093/bioinformatics/bty373.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Theres Leinonen (2010). "Gabmap - A Web Application For Dialectology". In:

Nerbonne, John and William A. Kretzschmar Jr. (2013). "Dialectometry++". In: *LLC* 28.1, pp. 2–12.

Neyman, Jerzy and Egon S. Pearson (1933). "On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231, pp. 289–337.

Niccoli, Matteo (2012). *The rainbow is dead... long live the rainbow! – Series outline*. https://mycarta.wordpress.com/2012/05/29/the-rainbow-is-dead-long-live-the-rainbow-series-outline/, (last accessed October 8, 2018).

Nilsson, Emelie and Anna-Karin Svensson (2004). *An ArcGIS Tutorial Concerning Transformations of Geographic Coordinate Systems, with a Concentration on the Systems Used in Lao PDR.*

Oesch, J. and A. Sidler (2017). *Videx: Indexing videos for the "official bulletin" of the swiss federal parliament*. www.jonasoesch.ch/portfolio/videx.

Parada, Carolina, Mark Dredze, Denis Filimonov, and Frederick Jelinek (2010). "Contextual Information Improves OOV Detection in Speech". In: *Proc. HLT*. HLT '10. Los Angeles, California: Association for Computational Linguistics, pp. 216–224. ISBN: 1-932432-65-5. URL: http://dl.acm.org/citation.cfm?id=1857999.1858024.

Paulo, Sérgio and Luís C. Oliveira (2004). "Automatic phonetic alignment and its confidence measures". In: *Proc. EsTAL*. Ed. by José Luis Vicedo, Patricio Martínez-Barco, Rafael Muńoz, and Maximiliano Saiz Noeda. Springer Berlin Heidelberg, pp. 36–44. URL: https://doi.org/10.1007/978-3-540-30228-5_4.

Pedersen, Carol and Joachim Diederich (2007). "Accent classification using support vector machines". In: *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*. IEEE, pp. 444–449.

Pellegrini, Thomas and Isabel Trancoso (Sept. 2010). "Improving ASR error detection with non-decoder based features". In: *Proc. Interspeech*. Makuhari, Chiba, Japan, pp. 1950–1953. URL: http://www.isca-speech.org/archive/interspeech_2010/i10_1950.html.

Pfister, Beat and Tobias Kaufmann (2008). "Sprachverarbeitung". In: *Grundlagen und Methoden der Sprachsynthese und Spracherkennung*.

Pickl, Simon and Jonas Rumpf (2012). "Dialectometric concepts of space: Towards a variant-based dialectometry". In: *Dialectological and Folk Dialectological Concepts of Space: Current Methods and Perspectives in Sociolinguistic Research on Dialect Change* 17, pp. 199–214.

Projekt Deutsch heute Orthografische Verschriftlichung gesprochener Sprache (Interviews und Map Tasks) KONVENTIONEN (2015). unpublished.

Quinlan, J. Ross (1993). *C4. 5: programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

— (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rahim, Mazin G., Chin-Hui Lee, and Biing-Hwang Juang (1997). "Discriminative utterance verification for connected digits recognition". In: *IEEE Transactions on Speech and Audio Processing* 5.3, pp. 266–277.

Ramers, Karl Heinz (1988). *Vokalquantität und -qualität im Deutschen*. Tübingen: Niemeyer.

Reichel, U. D. and T. Kisler (2014). "Language-independent grapheme-phoneme conversion and word stress assignment as a web service". In: *Elektronische Sprachverarbeitung 2014*. Ed. by R. Hoffmann. Vol. 71. Studientexte zur Sprachkommunikation. Dresden, Germany: TUDpress, pp. 42–49.

Reichel, Uwe D. (2012). "PermA and BALLOON: Tools for string alignment and text processing". In: *Proc. Interspeech*. Portland, Oregon, pp. 1874–1877.

— (2012–2018). *We talked about what improvements could be made to the proposed method using acoustic features to model regional variatoin.* Personal Communication. München, Germany.

Reichel, Uwe D., Florian Schiel, Thomas Kisler, and Nina Pörner (2016). "The BAS Speech Data Repository". In: *Proc. LREC*. Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1. URL: https://www.phonetik.uni-muenchen.de/forschung/publikationen/ReichelSchielKislerDraxlerPoerner-LREC2016.pdf.

Reubold, Ulrich and Jonathan Harrington (2017). "The influence of age on estimating sound change acoustically from longitudinal data". In: *Panel studies of variation and change*. Ed. by Suzanne Evans Wagner and Isabelle Buchstaller.

Rogowitz, B. E. and L. A. Treinish (1998). "Data visualization: the end of the rainbow". In: *IEEE Spectrum* 35.12, pp. 52–59. ISSN: 0018-9235. DOI: 10.1109/6.736450.

Rogowitz, Bernice E., Lloyd A. Treinish, and Steve Bryson (June 1996). "How Not to Lie with Visualization". In: *Computers in Physics* 10.3, pp. 268–273. ISSN: 0894-1866. DOI: 10.1063/1.4822401. URL: http://dx.doi.org/10.1063/1.4822401.

Rose, Richard C., Biing-Hwang Juang, and Chin-Hui Lee (1995). "A training procedure for verifying string hypotheses in continuous speech recognition". In: *Proc. ICASSP*. Vol. 1. IEEE, pp. 281–284.

Rowley, Anthony R. (1990). "The dialects of modern German: A linguistic survey." In: ed. by Charles Russ. London: Routledge. Chap. East Franconian, 394—416.

Rueber, Bernhard (1997). "Obtaining confidence measures from sentence probabilities." In: *Proc. Eurospeech*.

Russell, Stuart Jonathan, Peter Norvig, John F. Canny, Jitendra M. Malik, and Douglas D. Edwards (2010). *Artificial intelligence: a modern approach*. 3rd. Prentice hall Upper Saddle River.

Sato, Nobuo and Yasunari Obuchi (2007). "Emotion recognition using mel-frequency cepstral coefficients". In: *Information and Media Technologies* 2.3, pp. 835–848.

Schaaf, Thomas and Thomas Kemp (1997). "Confidence measures for spontaneous speech recognition". In: *Proc. ICASSP*. Vol. 3. IEEE, pp. 875–878.

Scherrer, Yves, Adrian Leemann, Marie-José Kolly, and Iwar Werlen (2012). "Dialäkt Äpp-A smartphone application for Swiss German dialects with great scientific potential". In: *Proc. SIDG*.

Scheutz, H. (1983). "Quantität und Lenis/Fortis im Mittelbairischen". In: *Beiträge zur bairischen und ostfränkischen Dialektologie. Ergebnisse der Zweiten Bairisch-Österreichischen Dialektologentagung*. Ed. by P. Wiesinger. Wien: Kümmerle Verlag, pp. 13 –33.

Schiel, Florian (1999). "Automatic Phonetic Transcription of Non-Prompted Speech". In: *Proc. ICPhS*. San Francisco, pp. 607–610.

— (2015). "A statistical model for predicting pronunciation". In: *Proc. ICPhS*. Glasgow, United Kingdom, paper 195.

Schiel, Florian and Thomas Kisler (2014). "German Alcohol Language Corpus - the Question of Dialect". In: *Proc. LREC*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 353–356. ISBN: 978-2-9517408-8-4.

Schmidt, Jürgen Erich, Joachim Herrgen, Tanja Giessler, Alfred Lameli, Alexandra Lenz, Karl-Heinz Müller, Wolfgang Näser, Jost Nickel, Roland Kehrein, Christoph Purschke, et al. (2001). *Digitaler Wenker-Atlas*. `http://www.diwa.info` (last accessed August 27, 2018).

Schmidt, Jürgen Erich, Joachim Herrgen, Roland Kehrein, Dennis Bock, Brigitte Ganswindt, Heiko Girnth, Slawomir Messner, Alfred Lameli, Christoph Purschke, and Anna Wolanska (2008). *Regionalsprache.de (REDE).–Forschungsplattform zu den modernen Regionalsprachen des Deutschen*.

Schmidt, Jürgen Erich and Joachim Herrgen (2011). *Sprachdynamik – Eine Einführung in die moderne Regionalsprachenforschung*. Berlin: Erich Schmidt Verlag.

Schuller, Björn, Stefan Steidl, and Anton Batliner (2009). "The INTERSPEECH 2009 emotion challenge". In: *Proc. Interspeech*.

Schuller, Björn, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan (2013). "Paralinguistics in speech and language—state-of-the-art and the challenge". In: *Computer Speech & Language* 27.1, pp. 4–39.

Schuller, Björn and Anton Batliner (2014). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, Björn W., Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss (2012). "The INTERSPEECH 2012 Speaker Trait Challenge." In: *Proc. Interspeech*, pp. 254–257.

Séguy, Jean (1973). *La dialectométrie dans l'Atlas linguistique de la Gascogne*. Société de linguistique romane.

Seigel, Mathew Stephen (2013). "Confidence Estimation for Automatic Speech Recognition Hypotheses". PhD thesis. University of Cambridge.

Seiler, Guido (2005). "On the development of the Bavarian quantity system". In: *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* 10.1, pp. 103–129.

Shen, Wade, Nancy Chen, and Douglas A Reynolds (2008). "Dialect recognition using adapted phonetic models". In: *Ninth Annual Conference of the International Speech Communication Association*.

Sinha, Shweta, Aruna Jain, and SS Agrawal (2015). "Acoustic-phonetic feature based dialect identification in Hindi speech." In: *International Journal on Smart Sensing & Intelligent Systems* 8.1.

SpeechDat(M): EU-project LRE-63314.

Stadtschnitzer, Michael, Christoph Schmidt, and Daniel Stein (2014). "Towards a Localised German Automatic Speech Recognition". In: *Proceedings of Speech Communication; 11. ITG Symposium*. VDE, pp. 1–3.

Stevens, Mary and Jonathan Harrington (2016). "The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English". In: *Journal of Phonetics* 58, pp. 118 –134. ISSN: 0095-4470. DOI: https://doi.org/10.1016/j.wocn.2016.08.003. URL: http://www.sciencedirect.com/science/article/pii/S0095447016300390.

Stoyanchev, S., P. Salletmayr, J. Yang, and J. Hirschberg (2012). "Localized detection of speech recognition errors". In: *Proc. SLT*, pp. 25–30. DOI: 10.1109/SLT.2012.6424164.

Sukkar, Rafid A., Anand R. Setlur, Chin-Hui Lee, and John Jacob (1997). "Verifying and correcting recognition string hypotheses using discriminative utterance verification". In: *Speech Communication* 22.4, pp. 333 –342. ISSN: 0167-6393. DOI: https://doi.org/10.1016/S0167-6393(97)00031-9. URL: http://www.sciencedirect.com/science/article/pii/S0167639397000319.

Tam, Y. C., Y. Lei, J. Zheng, and W. Wang (2014). "ASR error detection using recurrent neural network language model and complementary ASR". In: *Proc. ICASSP*, pp. 2312–2316. DOI: 10.1109/ICASSP.2014.6854012.

The ASR Consortium (1995). *Phondat2 Corpus (PD2)*.

Therneau, Terry and Beth Atkinson (2018). *rpart: Recursive Partitioning and Regression Trees*. URL: https://CRAN.R-project.org/package=rpart.

Tillmann, Hans Günther and Phil Mansell (1980). *Phonetik. Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Klett-Cotta.

Tiwari, Vibha (2010). "MFCC and its applications in speaker recognition". In: *International Journal on Emerging Technologies* 1.1, pp. 19–22.

Torgo, Luís, Rita P. Ribeiro, Bernhard Pfahringer, and Paula Branco (2013). "SMOTEfor Regression". In: *Proc. EPIA*. Ed. by Luís Correia, Luís Paulo Reis, and José Cascalho. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 378–389. ISBN: 978-3-642-40669-0. DOI: 10.1007/978-3-642-40669-0_33. URL: https://doi.org/10.1007/978-3-642-40669-0_33.

Trubetzkoy, Nikolai S. (1939). "Grundzüge der Phonologie (Traveaux du Cercle Linguistique de Prague, 7)". In: *Repr.(1968). Göttingen: Vandenhoek and Ruprecht*.

Vapnik, Vladimir (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media.

Vennemann, Theo (1991). "Certamen Phonologicum". In: ed. by Pier Marco et al.(eds.) Bertinetto. Torino: Rosebert & Sellier. Chap. Syllable structure and syllable cut prosodies in Modern Standard German, pp. 211–244.

Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2, pp. 260–269.

Wang, Lipo (2005). *Support vector machines: theory and applications*. Vol. 177. Studies in Fuzziness and Soft Computing. Springer Science & Business Media.

Watt, Dominic, Carmen Llamas, and Daniel Ezra Johnson (2014). "Sociolinguistic variation on the Scottish-English border". In: *Sociolinguistics in Scotland*. Springer, pp. 79–102.

Weintraub, Mitch, Francoise Beaufays, Zeév Rivlin, Yochai Konig, and Andreas Stolcke (1997). "Neural-network based measures of confidence for word recognition". In: *Proc. ICASSP*. Vol. 2. IEEE, pp. 887–890.

Wells, John C. et al. (1997). "SAMPA computer readable phonetic alphabet". In: *Handbook of standards and resources for spoken language systems* 4.

Wessel, Frank, Ralf Schluter, Klaus Macherey, and Hermann Ney (2001). "Confidence measures for large vocabulary continuous speech recognition". In: *IEEE Transactions on Speech and Audio Processing* 9.3, pp. 288–298.

Wieling, Martijn and John Nerbonne (2015). "Advances in dialectometry". In: *Linguistics* 1.

Wiese, Richard (2000). *The phonology of German*. Oxford University Press on Demand.

Wiesinger, Peter (1983). "Die Einteilung der deutschen Dialekte". In: *Dialektologie* 2. Halbband, pp. 807–900.

— (1990). "The Central and Southern Bavarian Dialects in Bavaria and Austria". In: *The dialects of modern German: A linguistic survey*. Ed. by Charles V. J. Russ, pp. 438–519.

William A. Kretzschmar, Jr (2006). "Art and Science in Computational Dialectology". In: *Special Issue on Progress in Dialectometry* 21.4, 12 pages.

Winkelmann, Raphael, Jonathan Harrington, and Klaus Jänsch (2017). "EMU-SDMS: Advanced speech database management and analysis in R". In: *Computer Speech & Language*.

Woehrling, Cécile and Philippe Boula de Mareüil (2006). "Identification of regional accents in French: perception and categorization". In: *Proc. Interspeech*.

Woehrling, Cécile, Philippe Boula de Mareüil, and Martine Adda-Decker (2009). "Linguistically-motivated automatic classification of regional French varieties." In: *Proc. Interspeech*, pp. 2183–2186.

Wolpert, David H. (1996). "The Lack of A Priori Distinctions Between Learning Algorithms". In: *Neural Computation* 8.7, pp. 1341–1390. ISSN: 0899-7667. DOI: 10.1162/neco.1996.8.7.1341.

Wong, E, J. Pelecanos, S. Myers, and S. Sridharan (2000). "Language identification using efficient Gaussian mixture model analysis". In: *Australian International Conference on Speech Science and Technology*. Vol. 4, pp. 7–6.

Woodland, P., C. Leggetter, James Odell, V. Valtchev, and Steve Young (Apr. 1998). "The Development Of The 1994 Htk Large Vocabulary Speech Recognition System". In: pp. 104–109.

Wrede, Ferdinand, Walther Mitzka, and Bernhard Martin (1927–1956). *Deutscher Sprachatlas. Auf Grund des von Georg Wenker begründeten Sprachatlas des Deutschen Reichs. 128 Karten*. Marburg: N. G. Elwert Verlag.

Wright, Marvin N and Andreas Ziegler (2015). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *arXiv preprint arXiv:1508.04409*.

Wängler, Hans-Heinrich (1967). *Grundriss einer Phonetik des Deutschen*. Marburg: N. G. Elwert Verlag.

Xue, Jian and Yunxin Zhao (2006). "Random Forests-Based Confidence Annotation Using Novel Features from Confusion Network". In: *Proc. ICASSP*. Vol. 1, pp. I–I. DOI: 10.1109/ICASSP.2006.1660229.

Young, Sheryl R. (1994). "Detecting misrecognitions and out-of-vocabulary words". In: *Proc. ICASSP*. Vol. 2. IEEE, pp. II–21.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. (2002). "The HTK book". In: *Cambridge university engineering department* 3, p. 175.

Zavareh, Farshid, Ingrid Zukerman, Su Nam Kim, and Thomas Kleinbauer (2013). "Error detection in automatic speech recognition". In: *Proc. ALTA*, pp. 101–105.

Zhang, Rong and Alexander I Rudnicky (2001). "Word level confidence annotation using combinations of features". In:

Zhou, Zhengyu and Helen M. Meng (2004). "A two-level schema for detecting recognition errors". In: *Proc. Interspeech*. Jeju Island, Korea, pp. 449–452. URL: http://www.isca-speech.org/archive/interspeech_2004/i04_0449.html.

Zhou, Zhengyu, Helen M. Meng, and Wai Kit Lo (2006). "A multi-pass error detection and correction framework for Mandarin LVCSR." In: *Proc. Interspeech*.

Zissman, Marc A. (1995). "Automatic Language Identification of Telephone Speech". In: *The Lincoln Laboratory Journal* 8.2, pp. 115–144.