Simon Wasserfall, BSc

# Automatic Speech Segmentation using Kaldi

## master's thesis

submitted to

**Graz University of Technology**

Supervisors

Dipl.-Ing. Dr.techn. Martin Hagmüller
Mag.rer.nat. Dr. Barbara Schuppler

**Signal Processing and Speech Communication Laboratory**

Graz, November, 2020

# Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

| | |
|---|---|
| date | (signature) |

## Abstract

Automatic speech segmentation is an often used method to annotate large speech corpora. It can serve as a starting point for corpus-based linguistic studies. In contrast to segmenting read speech, the segmentation of spontaneous, conversational speech is a more challenging task. Spontaneously pronounced words contain phenomena of reduction, assimilation and deletion and the task is therefore more complex than read speech. In this thesis, automatic speech segmentation is performed for the GRASS corpus, which contains both read and conversational speech data of Austrian German. The approach chosen for the segmentation is a forced alignment with the state of the art toolkit Kaldi. In addition to studying the impact of different frame-shifts during the acoustic modelling, also pronunciation modelling for Austrian German is a focus in this thesis. Pronunciation variation is modelled with a knowledge-based approach with the help of formalised phonological rules, resulting in a pronunciation lexicon. The results of a quantitative distance measure to reference alignments for the GRASS read speech component with 8.4%, is similar to previously reported values for the same speaking style. The analysis of the two speech style showed that the mean speechrate of conversational speech is more than twice as large as the mean speechrate of read speech.

## Kurzfassung

Automatische Segmentierung wird häufig dazu verwendet, umfangreiche Sprachdatensätze zu transkribieren. Die so generierten Annotationen bilden oft die Ausgangslage linguistischer Studien. In Spontansprache kommt es zu Phänomenen wie Auslöschungen, Angleichungen oder Substituierungen von einzelnen Lauten. Sie ist, im Vergleich zur gelesenen Sprache, für eine automatische Segmentierung eine besondere Herausforderung. In dieser Masterarbeit wird eine automatische Segmentierung des GRASS Korpusses durchgeführt. Dies beinhaltet sowohl die Segmentierung von gelesener Sprache als auch die Segmentierung von Spontansprache. Mit Hilfe des Kaldi Toolkits werden akustische Modelle trainiert und bekannte orthographischen Trankribierungen werden zu den entsprechenden Audiodaten ausgerichtet. Im Fokus der Arbeit steht die Untersuchung verschiedener frame-shifts, während die Audiomerkmale berechnet und verschiedene Varianten der Aussprache des Österreichischen Deutschs modelliert werden. Anhand eines wissensbasierten Ansatzes wird mit Hilfe von phonologischen Regeln ein Aussprachelexikon erstellt, welches die Ausprachevarianten des Korpusses abedeckt. Mit Hilfe eines Distansmaßes wird eine quantitative Evaluation zur gelesener Sprache durchgeführt. Die gesamte Distanz zwischen den automatischen Segmentierungen zu einem Referenzmaß ergibt einen Wert von 8.4%. Vorhergehende Studien berichten über ähnliche Distanzen bei der Evaluation von gelsener Sprache. Die Analyse der unterschiedlichen Sprachstile zeigt, dass die Sprechgeschwindigkeit während Spontansprache mehr als doppelt so hoch ist wie die Sprechgeschwindigkeit bei gelesener Sprache.

# Contents

# 1 Introduction

Automatic speech segmentation is a frequently used procedure to annotate large speech corpora (C. Van Bael et al., 2007). Segmentation on phone or word level of a corpus are the starting point for phonetic analysis and further linguistic research (Schuppler, Adda-Decker, and Morales-Cordovilla, 2014). Compared to manually created phonetic transcriptions, the usage of forced alignment promises significant time savings with the additional advantage of more consistent results. This thesis aims to do an automatic segmentation of the Graz corpus of read and spontaneous speech (GRASS) (Schuppler, Hagmüller, and Zahrer, 2017) with the help of the Kaldi Automatic Speech Recognition (ASR) toolkit (Povey et al., 2011).
There exist different forced alignment tools to align the orthographic transcriptions with the audio files on a phone or word level. One of these tools for German language is the MAUS tool (Schiel, 1999) that was used to segment some of the GRASS read speech (RS) component. The manually corrected results of the MAUS segmentation serve as a reference for the automatic segmentation with the Kaldi toolkit and a comparison gives an impression of the performance of the Kaldi based forced alignment procedure.
In contrast to speech material from RS, the conversational speaking style is a more challenging task for the forced alignment process (Bigi and Meunier, 2018). However, conversational speech (CS) is an important starting point for phonetic studies and therefore an automatic speech segmentation of the GRASS CS component is of interest. CS consists of more complex pronunciation variation than RS (Schuppler, Adda-Decker, and Morales-Cordovilla, 2014) and as a consequence, the pronunciation modelling during the forced alignment process is an important task.
A possibility to capture the pronunciation variation is the usage of a pronunciation lexicon that includes variants in addition to the canonical forms during the forced alignment process. With the help of existing studies on pronunciation variation of Austrian German, one can create a pronunciation lexicon that covers the pronunciation variation of the GRASS CS part. Automatic segmentation for the GRASS CS with the MAUS tool is not sufficient as the MAUS tool does not cover the pronunciation variation for Austrian German, so there is a need for performing a forced alignment specially for Austrian German, which can be achieved with the help of the open-source ASR toolkit Kaldi. In addition to studying the use of different frame-shifts during the acoustic modelling, also the pronunciation modelling for Austrian German is a focus in this thesis.

## 1.1 Forced alignment

In literature, *forced alignment* is a method for *automatic segmentation* with the help of a speech recognition system. The linguistic level of segmentation can be the *phonetic* level or the *word* level. *Forced alignment* describes a working mode of an ASR system, where the orthography is already given with the corresponding audio data. This section discusses the basic concepts of forced alignment and gives a summary of existing forced alignment systems, with a special focus on the characteristics of these alignment systems.

**Forced Alignment:**
In a forced alignment process, speech and its orthographic representation are aligned on a phone or word level. Figure 1.1 illustrates the typical input data for a forced alignment process. Pfister and Kaufmann (2017) point out that the orthographic text level differs from the speech level in many aspects. Text consists of a small amount of elements, the letters of the alphabet, and the resulting words are clearly separated. In spoken language, sounds are not clearly separated neither in temporal meaning nor in their characteristics. A discrete speech signal consists of a sequence of samples but there is no direct mapping between one sample and a corresponding sound. One step of the forced alignment is to find a mapping between the text level (grapheme level) and the phone level. The mapping is typically stored in a pronunciation lexicon and finding *pronunciation variants* is a core component to improve automatic phonetic segmentation through forced alignment. The second component of a forced aligner is a statistical model, also called acoustic model that models the realisations of phones. In a first step, the orthographic transcript of a speech utterance is mapped to the phone level resulting in a sequence of phones. With the sequence of phones and the speech input, the statistical model finds the best fitting alignment between the two given inputs. Figure 1.1 illustrates the forced alignment process with an example alignment of a spoken utterance of the GRASS RS component (Schuppler, Hagmüller, and Zahrer, 2017).

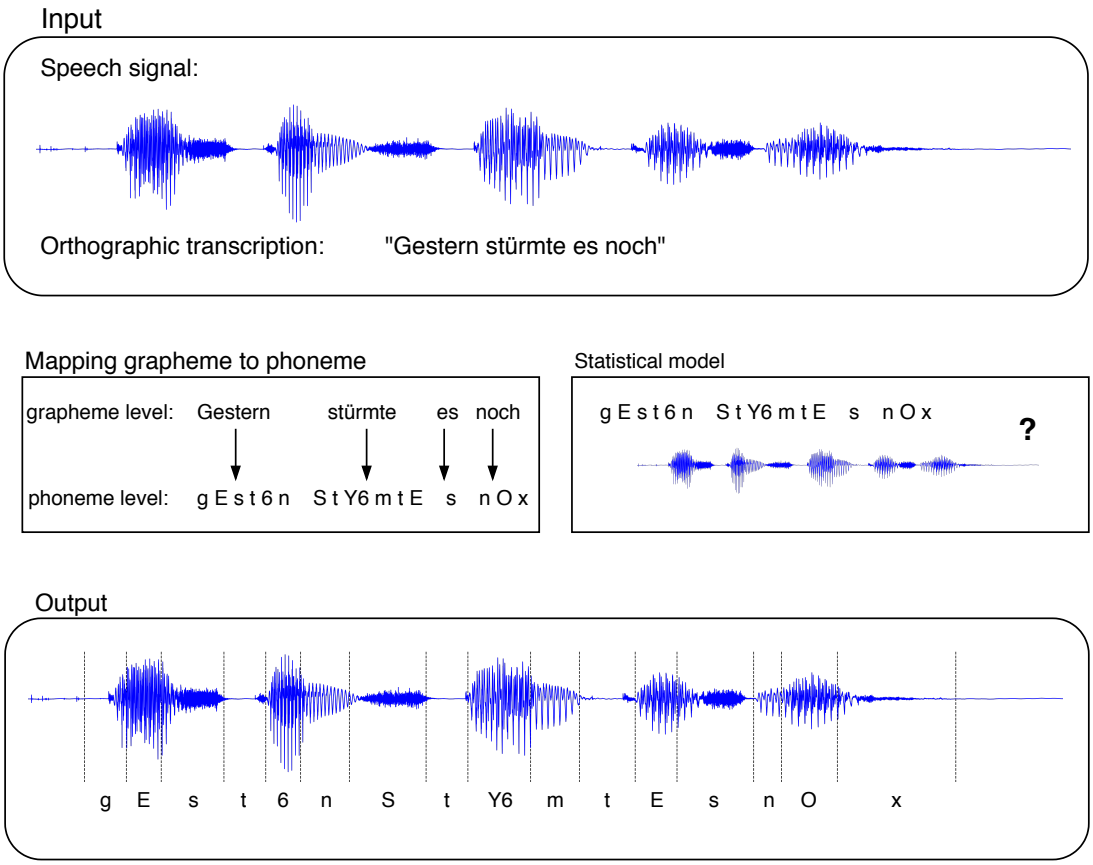


*Figure 1.1: Overview of the forced alignment process with an example from the GRASS RS component.*

**Architecture**
As mentioned in (McAuliffe et al., 2017), the architecture of the statistical model used in the forced alignment process is a characteristic to distinguish between various forced alignment systems. Most existing forced alignment systems use Hidden-Markov Models (HMM) as statistical model. Here are some points in which the acoustic models of forced alignment systems could

vary among each other:

- context dependent or context independent phone modelling (triphone vs. monophone models)
- account for speaker variability (e.g., acoustic feature transformations)
- topology of the HMM

**Training**
The same authors McAuliffe et al. (2017) point out that training the acoustic model can happen in distinct ways. Where some forced aligners work with *pre-trained* acoustic models of existing corpora, other forced aligners give the opportunity to *retrain* the acoustic model with the given data that should be aligned. The advantage of using existing corpora is that the acoustic models may be trained on manually created segmentations.

**Transcription constraining**
Johnson, Di Paolo, and Bell (2018) classified forced alignment in *unconstrained* alignment, where the aligner has less given information where to look for given words in the recording, whereas in *constrained* alignment, the transcription is time aligned to specific segments like utterances. Bigi and Meunier (2018) addressed a similar category for transcriptions that are approximated, which means that errors and omissions can occur during the annotation process.

**Toolkit**
There are different toolkits for performing an ASR task and each of them could be used to perform a forced alignment as well. Some of the open-source toolkits are listed below:

- HTK (S. J. Young and Sj Young, 1993)
- CMU Sphinx (Lamere et al., 2003)
- Kaldi (Povey et al., 2011)
- Julius (Lee, Kawahara, and Shikano, 2001)
- DeepSpeech (Hannun et al., 2014)
- RWTH ASR (Rybach et al., 2011)

One can compare characteristics regarding the structure of the toolkit like *used algorithms* or the *programming language* or one can compare characteristics regarding the usability of such systems like *preparation of the documentation*, *supported OS* or *license terms*. Because Kaldi supports state of the art algorithms within given template recipes, it was chosen for the automatic speech segmentation task in this thesis.

**Pronunciation modelling**
Pronunciation modelling is the task of dealing with different pronunciation variants. Especially for CS, pronunciation modelling is an important task to improve the automatic speech segmentation. Different concepts for pronunciation modelling within a speech recognition are summarised in (Chen et al., 2015). Figure 2.1 illustrates such concepts in a tree diagram.

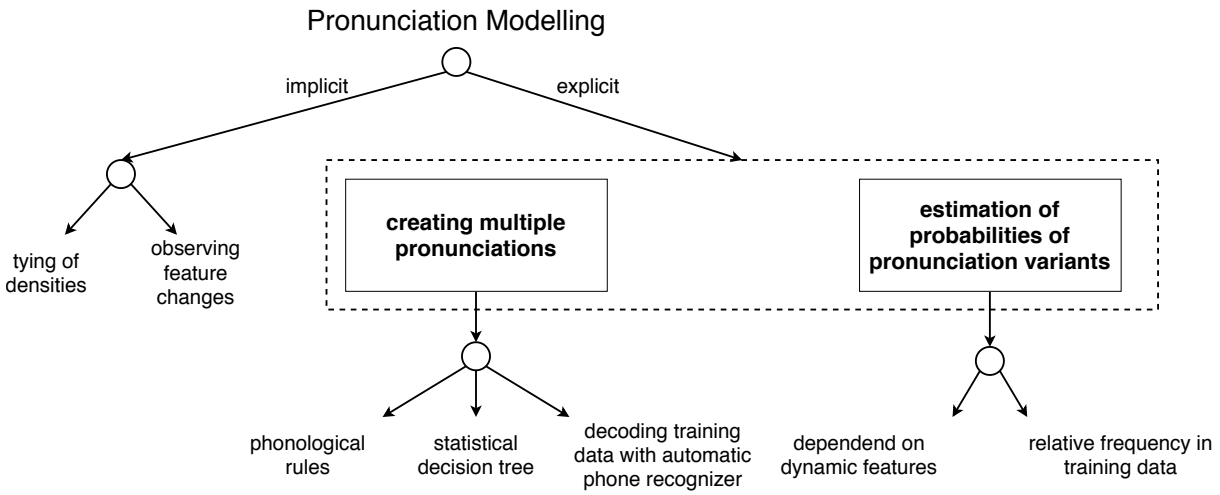


*Figure 1.2: Overview of pronunciation modelling within an ASR task.*

Each node in Figure 2.1 represents a decision criterion for designing a pronunciation model. By modelling the pronunciation variation within the acoustic model, the pronunciations are modelled in an *implicit* way (Hain, 2002). The *explicit* pronunciation modelling way consists of two parts. Schuppler, Adda-Decker, and Morales-Cordovilla (2014) created multiple pronunciation variants with a set of rules, which describe phonologically and phonetic reduction processes. By counting the frequencies of the pronunciation variants in the training data, one can estimate pronunciation probabilities and incorporate them into an ASR system (Chen et al., 2015).

**Forced alignment tools**

There are different forced alignment tools in order to perform an automatic speech segmentation task. Those tools are often based on the HTK toolkit. Table 1.1 lists alignment systems with some selected properties.

*Table 1.1: Some alignment systems and selected characteristics.*

| Alignment System | ASR Toolkit | Trainabilty on new data | Usage of pronunciation probabilities | Embedding within speech events |
|---|---|---|---|---|
| MAUS | HTK | ✗ | ✓ | ✗ |
| MFA | Kaldi | ✓ | ✓ | ✗ |
| SPPAS | Julius/HTK | ✓ | ✗ | ✓ |

The MAUS tool (Schiel, 1999), which is also available as a web-interface tool (Kisler, Uwe Reichel, and Schiel, 2017), works for more than 28 languages and also supports the German language with additional German dialect language settings. The MAUS tool uses a statistical pronunciation model, which also includes apriori probabilities. The Montreal Forced Alignment (MFA) tool (McAuliffe et al., 2017) is based on the Kaldi toolkit and supports the usage of triphone based acoustic models with an option of a speaker adaptive training. The trainability on new data allows to use an own corpus in order to create the acoustic model. The authors of the SPPAS tool (Bigi, 2012) addressed the problem of *within* speech, when performing an automatic speech segmentation task for spontaneous speech. The events filled pause, laughter, and noise are taken into account in the acoustic model for the French language (Bigi and Meunier, 2018).

# 2 Background

## 2.1 Pronunciation modelling

The main focus of this thesis is the pronunciation modelling during an automatic speech segmentation task for the GRASS corpus. This section gives an overview of the different methods existent for pronunciation modelling and is a summary of the paper by Strik and Cucchiarini (1999). A motivation for pronunciation modelling is the usage of CS corpora, as the amount of variation grows for CS compared to RS. A first distinction of pronunciation variation is the comparison between *interspeaker* and *intraspeaker* pronunciation variation. *Intraspeaker* variation describes the fact that one speaker can pronounce words in distinct ways depending on various factors such as *assimilation*, *co-articulation*, *reduction*, *deletion* or *insertion* phenomena. The degree of the mentioned factors depend on the speaking style, for example *reduction* phenomena are expected to be more present in CS than in RS. *Interspeaker* variation relates to the fact that words are pronounced differently depending on factors that describe differences between the speakers, like *region of origin*, *accents*, *sex* or *age*. Another factor for pronunciation variation is the *interlocutor*, which describes the speaker adapting his speech to the listener. The mentioned factors for pronunciation variation in speech can be summed up to *linguistic variation*, whereas there is also a variation caused by anatomical differences like vocal tract length or the variation caused by environmental influences (Lombard effect) (Vlaj and Kačič, 2011). In ASR systems, some of the mentioned pronunciation variation factors are solved by design, e.g. the temporal variation of phones with the usage of HMMs, whereas other factors can be modelled in an explicit way, like the usage of pronunciation lexicons. Nevertheless, it is not easy to say where pronunciation modelling begins and ASR improvements end, but it is a benefit to characterise pronunciation modelling in a form of a decision based framework.

**What type of pronunciation variation should be modelled?**
Mostly, pronunciation variation is modelled on a segmental level in contrast to variation depending on suprasegemental factors. On the segmental level one can distinguish between *word-internal* variation or *cross-word* variation processes. In an ASR task or forced alignment task, the *word-internal* variation can be modelled with the help of a pronunciation lexicon as the lexicon can be extended with additional pronunciation variants. To cover *cross-word* variation in a pronunciation lexicon, one can append *multi-word* entries that are entries with more than one word.

**Where should the information of the variation come from?**
In a *data-driven* approach, the information of the variation is derived from the data, whereas in a *knowledge* based approach the information is available in existing literature. Often both approaches are taken into account and the classification in a *data-driven* approach or *knowledge* based approach refer to the starting point of the research. To give an example, it is possible to start with a *knowledge* based approach in form of a pronunciation lexicon with variants and to analyse the pronunciation variants in a quantitative way with help of a corpus in a *data-driven* manner. A drawback of a *data-driven* approach is that the resulting information of the varia-

tion depends on the used data and the information can not be generalised to different situations. Within a *knowledge-based* approach a possible disadvantage could be the mismatch between the data and the information based on the literature. Another effect is overgeneralisation when the pronunciation lexicon contains variants that do not occur in the training data.

**Should the information be formalised or not?**
When formalising the information of variation one can use for instance rewrite rules (Schuppler, Adda-Decker, and Morales-Cordovilla, 2014) or artificial neural-networks (Deshmukh, Weber, and Picone, 1996) to obtain a more abstract representation of the information. Formalising the information has two main advantages. First, one has full control of the process of generating pronunciation variants and in addition, the information is present in an abstract representation. This makes it easier to adapt the process of pronunciation modelling to other corpora.

**In which component of the system should the variation be modelled?**
There are different components of a speech recognition system, where one can incorporate the knowledge of the variation. Similar to Strik and Cucchiarini (1999), the paper of Chen et al. (2015) summarises methods of pronunciation modelling and distinguishes between an *implicit* and an *explicit* pronunciation modelling. In an *implicit* way, one could model the information within the acoustic model. In some methods, acoustic parameters of a phone are tied together with similar phones in order to cover pronunciation variation. An *explicit* way to model pronunciation variation is to incorporate the knowledge in a pronunciation lexicon. There are different techniques of generating pronunciation variants. A *knowledge-based* approach is to use *phonological rules* to extend the word pronunciations in the lexicon. In a *data-driven* approach, one would decode labelled training data in order to obtain pronunciation alternatives of frequent words in the corpus. Figure 2.1 gives an overview of different methods for pronunciation modelling where each circle marks a decision criterion during the pronunciation modelling task.

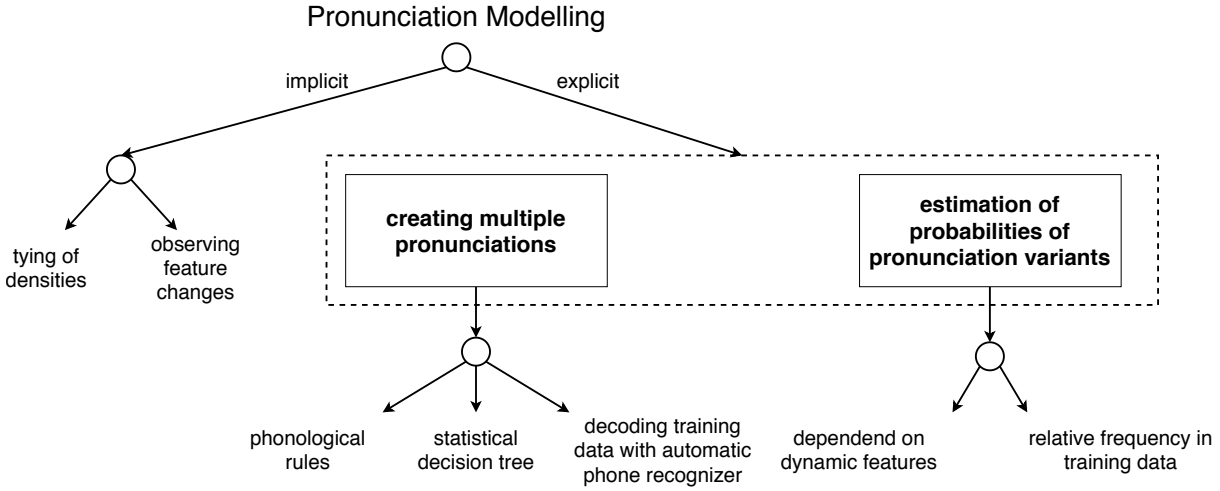


*Figure 2.1: Overview pronunciation modelling.*

The *knowledge-based* approach requires linguistic knowledge in the language and the speaking style of the speech corpora. In a study of Schuppler, Adda-Decker, and Morales-Cordovilla (2014), a *knowledge-based* approach is applied to the Austrian German GRASS corpus in order to obtain a broad phonetic transcription. In an *explicit* pronunciation model, the authors generate a pronunciation lexicon with the help of 32 phonological rules and perform a forced alignment task within HTK experiments. As this paper is the starting point for the pronunciation modelling in this thesis, said study is discussed in detail in Section 3.3.

## 2.2 Statistical description

As mentioned in the Introduction, the modelling of pronunciation is an important task for the forced alignment process and can be done in several ways. Schiel (2015) describes a statistical model for predicting pronunciation as an optimisation problem that contains an acoustic model and an apriori probability. The starting point is the statistical model for speech recognition and then I will derive the model for predicting pronunciations. Note that one can find an introduction to statistical speech recognition based on HMMs in (Pfister and Kaufmann, 2017) or (Gales and Steve Young, 2008).

In a first step, the input audio waveform is converted to a sequence of feature vectors $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$. Extracting information from the speech signal is explained in Section 2.3 feature extraction. The task of the speech recognition is to find a good estimation $\mathbf{W}$ of the word sequence $\mathbf{W} = \mathbf{w}_1\mathbf{w}_2...\mathbf{w}_K$ and with help of the maximum-a-posterior rule, the estimation problem can be written in a statistical way as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in V}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{X}), \tag{2.1}$$

where $V$ represents the set of the vocabulary. With the rule of Bayes it is straightforward to convert the above equation to

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in V}{\operatorname{argmax}} \frac{P(\mathbf{X}|\mathbf{W}) \cdot P(W)}{P(\mathbf{X})}, \tag{2.2}$$

where $P(\mathbf{X})$ denotes the acoustic evidence, which can be skipped as it is a scalar factor for the optimisation problem. Finally, the well-known formula can be written as

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in V}{\operatorname{argmax}} P(\mathbf{X}|\mathbf{W}) \cdot P(W), \tag{2.3}$$

where $P(\mathbf{X}|\mathbf{W})$ represents the *acoustic model* and $P(W)$ describes the *language model*. Since the acoustic model is not based on words, as a basic unit but on phones, each word $\mathbf{w}$ can be mapped into a sequence of base phones $\mathbf{k}^{\mathbf{w}}_{1:N} = k_1k_2,..k_N$, where $N$ is the number of used base phones. The likelihood can be computed with the help of the sum over all valid pronunciations with

$$P(\mathbf{X}|\mathbf{W}) = \sum_{\mathbf{K}} P(\mathbf{X}|\mathbf{K})P(\mathbf{K}|\mathbf{W}), \tag{2.4}$$

where $\mathbf{K}$ is a particular sequence of pronunciations. The conditional probability that the pronunciation sequence $\mathbf{K}$ occurs for a given word sequence $\mathbf{W}$ can be expressed by multiplying all pronunciation probabilities for each word in the word sequence:

$$P(\mathbf{K}|\mathbf{W}) = \prod_{l=1}^{L} P(\mathbf{k}^{\mathbf{w}_l}|\mathbf{w}_l). \tag{2.5}$$

In a recognition task, one can approximate the sum of Equation 2.4 into an optimisation

problem

$$P(\mathbf{X}|\mathbf{W}) = \underset{\mathbf{K}\in\Psi}{\operatorname{argmax}}\, P(\mathbf{X}|\mathbf{K})P(\mathbf{K}|\mathbf{W}), \tag{2.6}$$

where $\Psi$ is the space of possible pronunciations for a given word sequence $\mathbf{W}$ and depends on the pronunciation lexicon. By formulating Equation 2.5 as an optimisation problem, the recogniser treats pronunciation variants as alternative word hypotheses and by inserting Equation 2.6 in Equation 2.3, one can see that the recognition tasks consists of two different optimisation steps. It is now possible to relate Equation 2.6 to the described statistical model for predicting pronunciation in the paper of Schiel (2015) and therefore to a forced alignment task. In contrast to a speech recognition task, the objective of a forced alignment is not to estimate the best word sequence $\hat{\mathbf{W}}$, like in 2.3, but to estimate the best fitting pronunciation sequence. As a consequence, the two optimisation problems are simplified to the following optimisation problem:

$$\hat{\mathbf{K}} = \underset{\mathbf{K}\in\Psi}{\operatorname{argmax}}\, P(\mathbf{X}|\mathbf{K})P(\mathbf{K}|\mathbf{W}), \tag{2.7}$$

where $\hat{\mathbf{K}}$ describes the optimal pronunciation sequence in the search space $\Psi$. Schiel (2015) states that automatic segmentation and labelling systems mainly differ in the search space $\Psi$ and the way how the apriori probability P($\mathbf{K}|\mathbf{W}$) is incorporated into the system.

## 2.3 Feature extraction

Dave (2013) described different feature extraction techniques for a speech recognition task. This section focuses on the Mel Frequency Cepstral Coefficients (MFCC), as they are used for the forced alignment task of this thesis. With the help of the cepstrum, one can decompose the speech signal production mechanism and the mechanism of sound shaping within the vocal tract. In the quefrency domain the lower coefficients represent the characteristics of the vocal tract with its specific shape and resonance frequencies and the higher quefrency coefficients represent the speech excitation signal. The first coefficient $c(0)$ relates to the logarithm of the signal energy. The preprocessing step consists of a non-uniformly spaced mel-filterbank with which the human audio perception is taken into account. The procedure of calculating the MFCC features is shown in Figure 2.2.

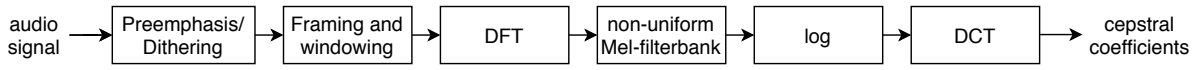


*Figure 2.2: Calculating the cepstral coefficients during feature extraction.*

In a preprocessing step, a preemphasis filter and dithering is applied to the audio signal. Afterwards the audio signal is chunked into frames, which are multiplied by a window function in order to reduce boundary effects. With the help of the Fast Fourier Transformation (FFT), the Discrete Fourier Transformation (DFT) is implemented in an efficient way. In a next step the power spectrum is calculated in the frequency domain and, by applying the triangular filters of the Mel-filterbank, the energy in each mel-bin is calculated. After converting to the logarithmic domain and applying the Discrete Cosinte Transform (DCT), one obtains the cepstral coefficients. Pfister and Kaufmann (2017) describe the MFCC extraction in a more detailed way.

Important parameters during a feature extraction process are the *frame length* and the *frame*

*shift*, the *number of mel bands* and the *number of cepstral coefficients*. The default *frame shift* in the Kaldi toolkit is 10 ms. Hämäläinen et al. (2009) uses a *frame shift* of 5 ms in order to allow the processing of smaller speech segments. This thesis investigates the impact of different *frame shifts* on the automatic speech segmentation by using *frame shifts* of 10 ms, 7 ms and 5 ms.

## 2.4 Acoustic modelling

Section 2.2 introduces the *acoustic* model with the conditional probability $P(\mathbf{X}|\mathbf{W})$ with the sequence of feature vectors $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2...\mathbf{x}_T$ and the sequence of words $\mathbf{W} = \mathbf{w}_1\mathbf{w}_2...\mathbf{w}_K$. With the help of the pronunciation lexicon, one can convert the sequence of words into a sequence of base phones where one specific sequence is called $\mathbf{K}$. A description of the acoustic model is the *Hidden Markov Model* (HMM), which consists of two coupled statistical processes. The internal process is the markov model with a specific number of states and transitions. Emitting states are coupled with observation distributions that give observation probabilities for a given feature vector. Because the feature vector has multiple dimensions also the observation distributions are multi-dimensional and, as each cepstral coefficient is set of the real numbers, the resulting HMM is called *continuous density* HMM. Figure 2.3 illustrates the topology of a *Kaldi* HMM for a non-silence phone,

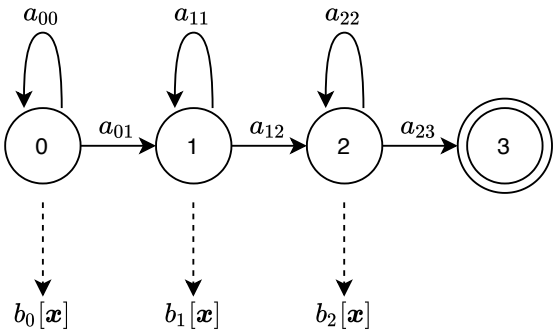


*Figure 2.3: HMM topology of a non-silence phone.*

In Figure 2.3, $a_{ij}$ represents the transition probabilities between the states and $b_j[\mathbf{x}]$ represents the observation distributions for a specific state $j$. Note that the final state 3 does not emit an observation distribution and is called *non-emitting* state. In Figure 2.3, each state only depends on the previous state and each observation only depends on the current state. This kind of HMM is called a *linear* HMM. For the description of multivariate continuous distributions, the *Gaussian Mixture Model* (GMM) is a useful concept and

$$b_j[\mathbf{x}] = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{x}; \mu_{jm}, \Sigma_{jm}) \tag{2.8}$$

represents an observation distribution for a specific state $j$, with a sum over all components $m$ and the prior probabilty $c_{jm}$.

Gales and Steve Young (2008) described the architecture of a HMM-based speech recognition system and the various refinement techniques in order to achieve state-of-the-art performance. A starting point in the Kaldi toolkit is a *monophone* based HMM acoustic model, which is a context independent phone model. A context dependent phone model is the *triphone* model,

which can be achieved by refining the monophone acoustic model. McAuliffe et al. (2017) investigated the impact of a monophone acoustic model in contrast to a a triphone based acoustic model with speaker-adapted features on the forced alignment process. Within this thesis, a monophone acoustic model is used to see the explicit pronunciation modelling effects.

# 3 Materials & Methods

## 3.1 GRASS corpus

The *Graz corpus of read and spontaneous speech* (GRASS) (Schuppler, Hagmüller, and Zahrer, 2017) was recorded in the soundproof recording studio of the SPSC laboratory of TU Graz. It is a large scale speech database for Austrian German of approximately 30 hours of read and spontaneous speech from 38 speakers. The corpus was designed to be suitable for linguistic and phonetic studies as well as serving as the starting point for speech recognition systems. The corpus promises high-quality super wideband recordings in order to simulate different acoustic environments by convolution methods. The material for the RS component are phonetically balanced sentences and digits from each speaker. The material for the CS component consists of 19 conversations with a length of approximately one hour each, allowing to model pronunciation variation and other speaker phenomena of spontaneous speech. The high quality orthographic transcriptions allow further automatic procedures in order to create different annotation layers.

**Read speech**
Each of the 38 speakers read approximately 62 phonetically balanced sentences, which were motivated by the Kiel corpus, and four telephone numbers as well. In addition, the speakers read 10 conversational like sentences in order to get a point of intersection to the CS component. In total, the RS component contains 19 511 word tokens from 1 660 word types and 2 774 different utterances.

**Conversational speech**
In order to reduce the dialectal variation in the CS component, all speakers were born and grew up in Austria and are currently living in Graz or Vienna, and had at least high school degrees. Additional information e.g., education level, region of childhood, working area, etc., were collected from the speakers for possible further linguistic or phonetic studies. The material of the CS component consists of 19 conversations with mixed pairs and gender-homogenous pairs with a length of approximately one hour each. The annotations of the CS component are done by six linguistically educated transcribers in a supervised manner and an additional third independent annotator in order to obtain high quality orthographic transcriptions. In order to do an automatic phonetic speech segmentation, the speech was segmented into chunks of maximal four seconds length. Furthermore, detailed annotations as for instance laughter, backchannels, broken words, foreign words, or noise were annotated. The detailed annotations helped to create an accurate lexicon for automatic speech segmentation or automatic speech recognition tasks.

## 3.2 Kaldi overview

Kaldi (Povey et al., 2011) is an open-source speech recognition toolkit, which is frequently used by researchers in the field of speech recognition. Its goal is to provide a modular code structure that is easy to understand and extend. The core components or used libraries are the

*OpenFST* toolkit to allow the implementation of finite state transducers and the *BLAS* and *LAPACK* libraries to support linear algebra routines. On top of the libraries, Kaldi classes are implemented in C++ and the user can easily interact with the Kaldi toolkit in form of existing shell scripts that are designed to have one specific functionality. Figure 3.1 illustrates the structure of the Kaldi toolkit and shows some of the implemented C++ classes.

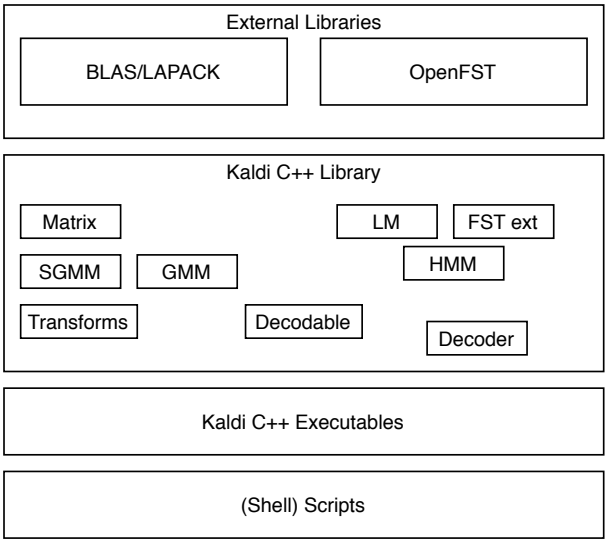


*Figure 3.1: Kaldi structure overview motivated by (Povey et al., 2011).*

Weighted finite-state transducers (WFST) (Mohri, Pereira, and Riley, 2008) provide a common and natural representation for major components of an ASR system, such as *HMMs*, *context-dependency models*, *pronunciation dicitionaries*, *statistical grammar models*, and *word or phone lattices*. Finite-state transducers are closely related to finite automatons, whose state transitions are labelled with both input and output symbols. When adding weights to the finite-state transducers one can encode probabilities, durations, penalties, or other quantities that accumulate along a path through the transducer. The usage of WFSTs for a speech recognition system promises an efficient implementation with new optimisation opportunities by implementing each part of an ASR system as a separate transducer and combining them afterwards. In the Kaldi toolkit, following transducers are implemented and combined to an overall transducer for the training and decoding step:

- G → encodes the grammar or language model
- L → represents the lexicon: output symbols are words and input symbols are phones
- C → represents the context-dependencies: its output symbols are phones and its input symbols represent context-dependent phones (windows of N phones)
- H → contains the HMM definitions: output symbols represent context-dependent phones and its input symbols are transition IDs

This thesis focuses on pronunciation modelling during a forced alignment task, which can be achieved with a pronunciation lexicon. Figure 3.2 shows a basic lexicon FST for a lexicon with the two entries: noch [n o x] and ja [j a]. The symbols above the arches between two states indicate *input symbol* : *output symbol*/*weight* and the state 0/0 indicates a *finite state*.

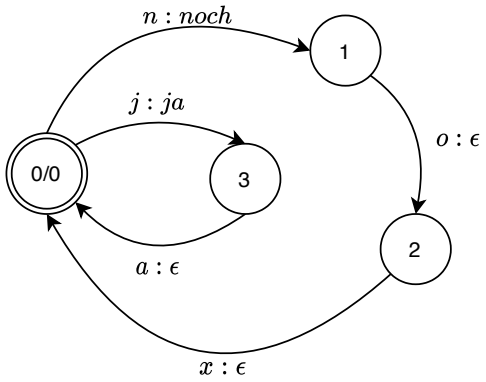


*Figure 3.2: Basic lexicon FST model for the two word entries noch [n o x] and ja [j a]. Note that ϵ represents an empty output symbol.*

## 3.3 Pronunciation modelling for Austrian German

Section 2.1 summarised different methods to perform pronunciation modelling and showed that one can model the pronunciations in an *implicit* way in the acoustic model or in an *explicit* way in form of a pronunciation lexicon. Figure 3.3 illustrates the approach chosen for modelling variants typical for Austrian German and for CS in all German varieties. First, pronunciation variants were created, then their probabilities were estimated by their relative frequencies in the training data. For the CS component, annotations in form of phonetic segmentations do not exist and as a consequence the pronunciation modelling was done with a knowledge-based approach. Starting point for the pronunciation modelling was the work of Schuppler, Adda-Decker, and Morales-Cordovilla (2014), in which pronunciation modelling was performed with the help of phonological rules based on existing linguistic studies on Austrian German. A part of the phonological rules are also valid for spoken language of German spoken in Germany and a part of the phonological rules are specific for Austrian German. The next sections describe the creation of pronunciation variants for the RS and the CS component of the GRASS corpus.

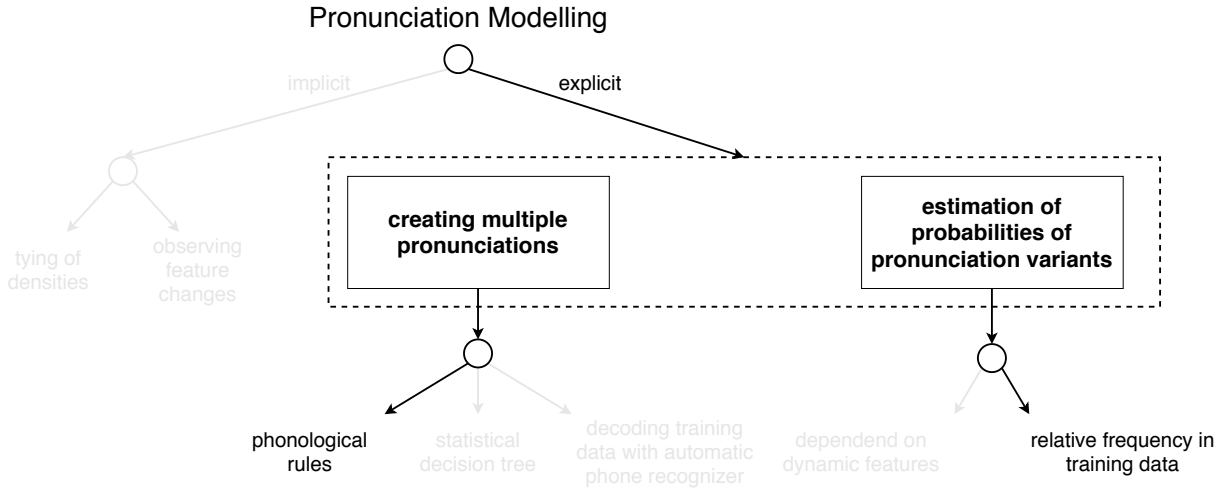


*Figure 3.3: Solution approach for the knowledge-based pronunciation modelling for Austrian German.*

### 3.3.1 Creating pronunciation variants for Austrian German

This section describes the creation of the lexicon with pronunciation variants, which is similar to the method presented by Schuppler, Adda-Decker, and Morales-Cordovilla (2014). Figure 3.4 illustrates the most important steps I took to obtain the final lexicon with pronunciation variants.

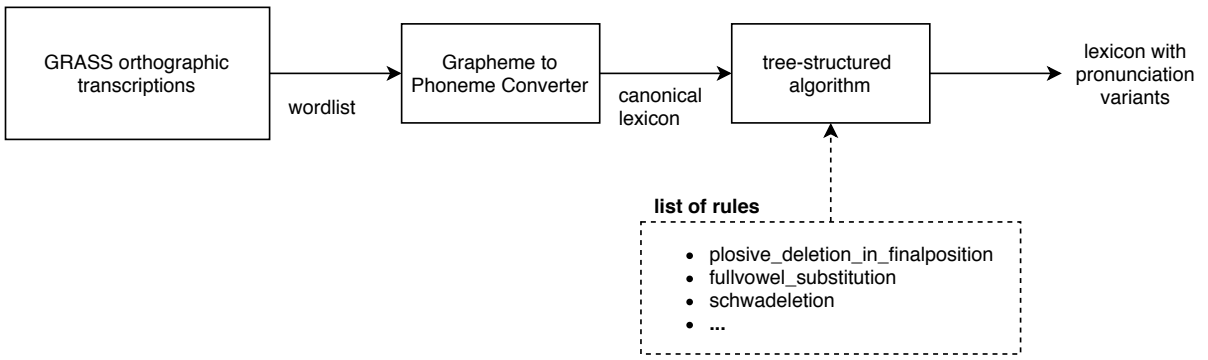


*Figure 3.4: Overview generating pronunciation variants.*

In a first step, a unique word list was created from all orthographic transcriptions. The online Tool G2P (U.D. Reichel, 2012) converted the words into canonical word pronunciations and Figure 3.5 shows an example output of the G2P tool:

```
AUFDREHEN;? ' aU f . d r e: . @ n
BLAUEN;b l ' aU . @ n
DIGITALES;d i . g i . t ' a: . l @ s
```

*Figure 3.5: Example output of the G2P tool with stress and syllabic information.*

The next procedure explains the augmentation of the raw lexicon to a lexicon with pronunciation variants. After obtaining a lexicon with canonical pronunciations for each occurring word, a tree-structured algorithm generated pronunciation variants with the help of phonological rules. There were 16 phonological rules reflecting assimilation and deletions that are typical for all varieties of spoken German and there are 13 phonological rules that reflect processes typically for Austrian German. An overview of the used phonological rules with an example is given in Table 3.1. Some of the rules use information about the syllabic structure and word stress (Schuppler, Adda-Decker, and Morales-Cordovilla, 2014). The rules were applied in a tree-structured algorithm, i.e., that previous generated pronunciation variants were used for the next step as well as the canonical pronunciation variant. The advantage of a tree-structured algorithm is that, compared to other algorithms, the number of pronunciation variants is large.

*Table 3.1: Overview of phonological rules used for pronunciation modelling of Austrian German.*

| | **# phonological rules** | **# deletion rules** | **# substitution rules** | **Example** |
|---|---|---|---|---|
| Assimilation and deletions typical for all varieties of spoken German | 16 | 11 | 5 | Deletion of schwa: a n m a x @ n → a n m a x n |
| Assimilation and deletions typical for Austrian German | 13 | 4 | 9 | Full vowel substitution: a p l aI t @ n→ o p l aI t @ n |

### 3.3.2 Pronunciation lexicon read speech

For the RS component of the GRASS corpus, I created three different lexicons with different numbers of phonological rules. Similar to Adda-Decker and Lamel (1999), the complexity of a lexicon is defined by the ratio between the *total number of variants* and the *total number of (canonical) entries* in the lexicon. Table 3.2 gives an overview of the different lexicons for the RS component. Starting point for the creation of the pronunciation lexicon with variants was a word list from the orthographic transcriptions of the RS component with 2 059 word types.

*Table 3.2: Overview of the used lexicons for the GRASS RS component.*

| **lexicon name** | **# phonological rules (#deletion rules/#substitution rules)** | **average # variants (complexity)** | **number of phones** |
|---|---|---|---|
| canonical Austrian German | 2 (D 0/S 2) | 0 | 60 |
| read speech rules | 8 (D 4/S 4) | 0.64 | 60 |
| overgeneralised + canonical Austrian German | 27 (D 14/S 13) | 3.40 | 60 |

Note that for all lexicons, two phonological rules were implemented as replacement rules, i.e., that the converted pronunciation was replaced in the pronunciation lexicon instead of added as an additional pronunciation variant. The two phonological rules define the *canonical Austrian German* lexicon. The first rule was the word-final -ig realisation as /ik/ and the second rule was the devoicing of the alveolar fricative /z/.

### 3.3.3 Pronunciation lexicon conversational speech

For the *GRASS* CS component, there were transcriptions details in form of tags. Some of the tags marked following words as *dialect words*, *foreign language words*, or words that are not spoken entirely, also called *broken words*. With the information from the transcription, one could create separated lexicons, which allowed to also have separated processing strategies. Table 3.3 gives an overview of the different lexicons used for the *GRASS* CS component and it shows whether the G2P tool was used.

*Table 3.3: Different lexicons for GRASS CS processing.*

| **lexicon name** | **description** | **usage G2P** | **G2P language setting** |
|---|---|---|---|
| foreign words | lexicon with words which were marked as foreign words in the transcription | ✓ | eng, deu, spa-ES, fra-FR, swe-SE, ita-IT |
| dialect words | lexicon with words which were marked as dialect words in the transcription | ✓ | deu |
| broken words | lexicon with words which were marked as broken words in the transcription | ✓ | deu |
| Austrian German variants | lexicon with pronunciation variants, generated with phonological rules | ✓ | deu |
| Austrian German manually | lexicon with special Austrian German words, created manually | ✗ | ✗ |

For the lexicon `foreign words`, the G2P tool was used with different language settings than the German language setting *deu*. The resulting pronunciations consisted of a different phone set than the German phone set. With a phone mapping, one could make sure that the phone set was not extended. The processing structure of the different lexicons is illustrated in Figure 3.6 and starting point for the lexicons `foreign words`, `dialect words`, `broken words` and `Austrian German variants` was a word list of 12 925 word types, which was gathered from all occurring words in the transcription files of the GRASS CS component. For some lexicons, it was necessary to have a phone mapping to ensure the phone set of 60 phones.

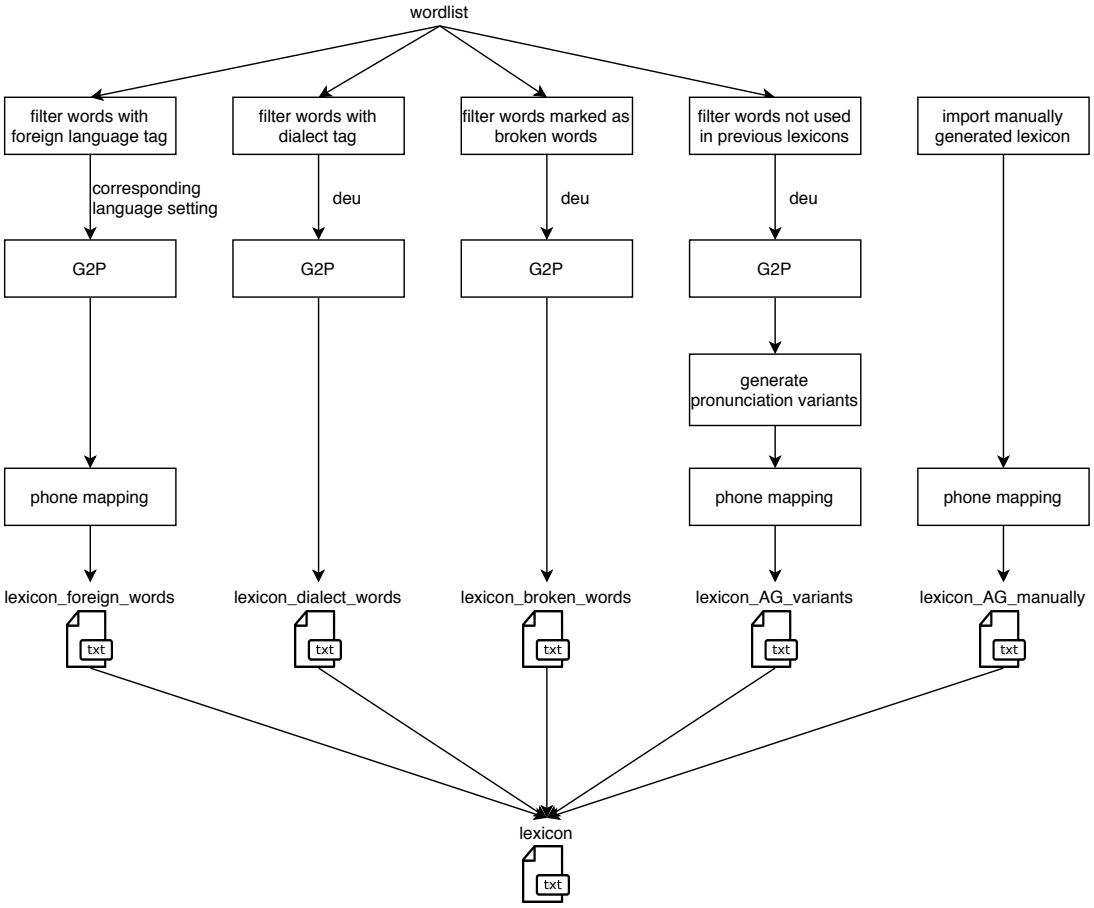


*Figure 3.6: Overview generating lexicons for GRASS CS component.*

The following paragraphs summarise information about the distinct lexicons.

**Lexicon foreign words**
The lexicon of foreign words consisted of 641 word types and 9 different languages other than German. Most of the foreign words were in English. Table C.2 in Appendix C shows the mapping between the tag of the transcription and the G2P language setting. Note that not all occurring languages were supported from the G2P tool, so a different language setting was used instead. In order to contain a similar phone set, a phone mapping was applied after the usage of the G2P tool. Table C.1 in the Appendix C illustrates the phone mapping of the foreign language words to an Austrian German phone set.

**Lexicon dialect words**
With the help of the tagged dialect words, it was possible to filter out these words from the word list and the lexicon consists of 306 word types. The canonical pronunciation of the dialect words were obtained with the G2P tool and the German language setting. There was no need to create pronunciation variants for theses words, as their orthography already reflects their pronunciation.

**Lexicon broken words**
In CS, a common phenomenon is that words are not pronounced entirely but just their beginning, i.e., broken word. Fortunately, these words were marked in the orthographic transcription, thus they were filtered out (618 word types) from the word list. The G2P tool produced a systematical error when the broken word did not contain any vowel. A manual correction solved these errors after the G2P output.

**Lexicon Austrian German with pronunciation variants**
Similar to the procedure described for the RS component, the pronunciation variants were generated with the help of the phonological rules. Depending on the rule set, there were different lexicons with a different complexity. Some of the words were automatically detected as foreign words from the G2P tool. Therefore a similar phone mapping than for the foreign word lexicon was applied after the generation of the pronunciation variants. The mapping details are provided in Appendix C.

**Lexicon Austrian German manually**
Some Austrian German pronunciation variants could not be generated by automated rules, and thus there was a need for importing a manually generated lexicon for frequently used words with a specific pronunciation. The resulting lexicon contained 83 word types.

Finally, all lexicons were joined to a combined lexicon used for the FA task. Depending on the used phonological rules, there were three lexicons for the CS component. An overview of the lexicons and the used rules is shown in Table 3.4.

*Table 3.4: Overview of the used lexicons for the GRASS CS component.*

| lexicon name | # phonological rules (#deletion rules/#substitution rules) | average # variants (complexity) | number of phones |
|---|---|---|---|
| deletion rules | 22 (D 15/S 7) | 2.98 | 63 |
| overgeneralised | 29 (D 15/S 14) | 4.73 | 63 |
| overgeneralised + canonical Austrian German | 29 (D 15/S 14) | 4.09 | 60 |

### 3.3.4 Estimation of pronunciation probabilities

As mentioned in Section 3.3, the second part of explicit pronunciation modelling is the estimation of the pronunciation probabilities. A straightforward and common way to estimate the pronunciation probabilities is to determine the relative frequency of the pronunciations in the training data. Similar to the work by Chen et al. (2015), I formulate the pronunciation probability as

$$\pi(\mathbf{k}_i^{\mathbf{w}}|\mathbf{w}) = \frac{C(\mathbf{w}, \mathbf{k}_i^{\mathbf{w}}) + \lambda_1}{\sum_{i=1}^{N_w}(C(\mathbf{w}, \mathbf{k}_i^{\mathbf{w}}) + \lambda_1)}, \tag{3.1}$$

where $\mathbf{k}_i^{\mathbf{w}}$ is the $i^{th}$ pronunciation of the word $\mathbf{w}$ with $N_w$ different word pronunciations. The count of the word pronunciation is denoted as $C(\mathbf{w},\mathbf{k}_i^{\mathbf{w}})$ and $\lambda_1$ is a smoothing constant, typically set to 1.

## 3.4 Acoustic models

This section describes the acoustic models that were used during the forced alignment with Kaldi. The focus of the forced alignment process is to model different frame-shifts during the feature extraction and to use pronunciation lexicons with different complexities. Table 3.5 lists the most important feature extraction settings, with which the feature extraction was calculated for the forced alignment task.

*Table 3.5: Part of feature extraction settings in Kaldi.*

| parameter | value |
|---|---|
| frame-length | 25 ms, 30 ms |
| frame-shift | 5 ms, 7 ms, 10 ms |
| # melfilterbank bins | 23 |
| # cepstral coefficients | 13 |
| window type | 'povey' |
| sample-frequency | 48 kHz |

**Read speech**
The acoustic model for RS was a monophone acoustic model with GMM as observation densities. The topology for the 60 non-silence phones consisted of a linear continuous density HMM with three states and an additional fourth, non-emitting, final, state. The training data of the acoustic monophone model was the entire RS component with 4 449 utterances spoken from 38 different speakers. There was no development or evaluation set, as it is a common approach to use all training data for the forced alignment process (McAuliffe et al., 2017). A cepstral mean voice tract normalisation (CMVTN) was applied after the feature extraction. The training procedure of the acoustic model was done in an iterative way with the *Viterbi* algorithm. The training of the models was calculated with different frame-shifts and with three lexicons containing different complexities. For details about the lexicons see Section 3.3.

**Conversational speech**
In analogy to the acoustic model of RS, the CS acoustic model was also a monophone model with GMM as observation densities. The training data for the CS component consisted of 19 conversations from 38 speakers resulting in approximately 19 hours of speech. In addition to the monophone model, I tested a triphone based acoustic model with a Subspace Gaussian Mixture Models (SGMM) for the observation densities. The motivation for this acoustic model is the fact that this model leads to the best WER for an ASR task as found in experiments for a baseline ASR model for the GRASS corpus.

## 3.5 Workflow

The workflow for performing the task of forced alignment, shown in Figure 3.7, consists of three major steps. In the *Preprocessing* step the data from the *GRASS* corpus is prepared for the Kaldi required input format. Once the data is prepared for *Kaldi*, I applied different tasks within Kaldi. In a first step, an acoustic model was trained and the data was aligned to the model afterwards. In the *Postprocessing* step, the alignments from Kaldi were imported to the *Python* domain in a user-friendly readable form. After importing the Kaldi alignments, I performed different tasks like calculating statistics of the used pronunciation variants, comparing different alignments or exporting the alignments to TextGrid files in order to visualise them in the *Praat* software (Boersma and Weenink, 2018). Note that the preprocessing and the postprocessing code is written in *Python* and Kaldi tasks were performed with *bash-scripts*.

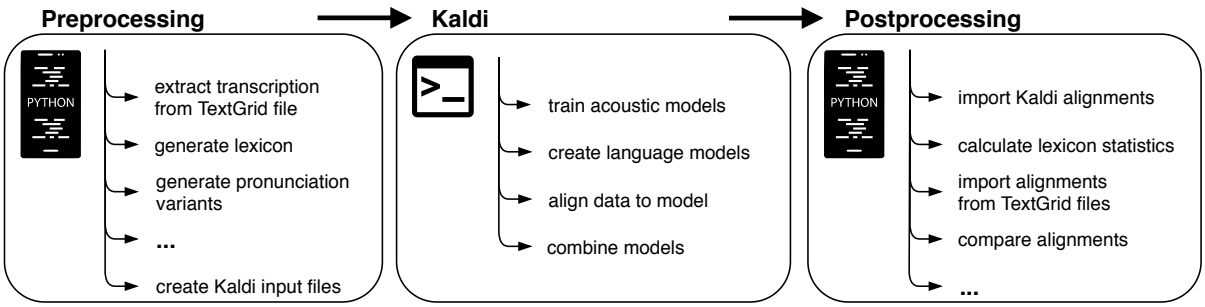


*Figure 3.7: Structure of the working flow for forced alignment using Kaldi.*

### Preprocessing

In the preprocessing step, the data from the GRASS corpus was prepared to fit the Kaldi specific input format. In the class `KaldiPreProcessor`, the main text files necessary for Kaldi were prepared. Note that there is a different class for the RS component than for the CS component, because the information of the annotations and the audio data are in a different format. In the RS component, there is one audio file per utterance, whereas in the CS component there are long audio files with multiple spoken utterances or chunks and a corresponding Praat TextGrid file, with the information about the segmentation of the chunks. After importing the transcription information, all occurring words are gathered in a word list, which is the starting point for generating the lexicon. The class `LexiconGenerator` takes the word list as an input and generates the lexicon with Austrian German variants and, in the case of the GRASS CS component, other lexicons. Inside the class `PronVarGenerator` the phonological rules are implemented and, depending on the user input, specific rules are selected and applied in order to generate the lexicon with pronunciation variants. Figure 3.8 illustrates the folder structure of one *Kaldi recipe*.

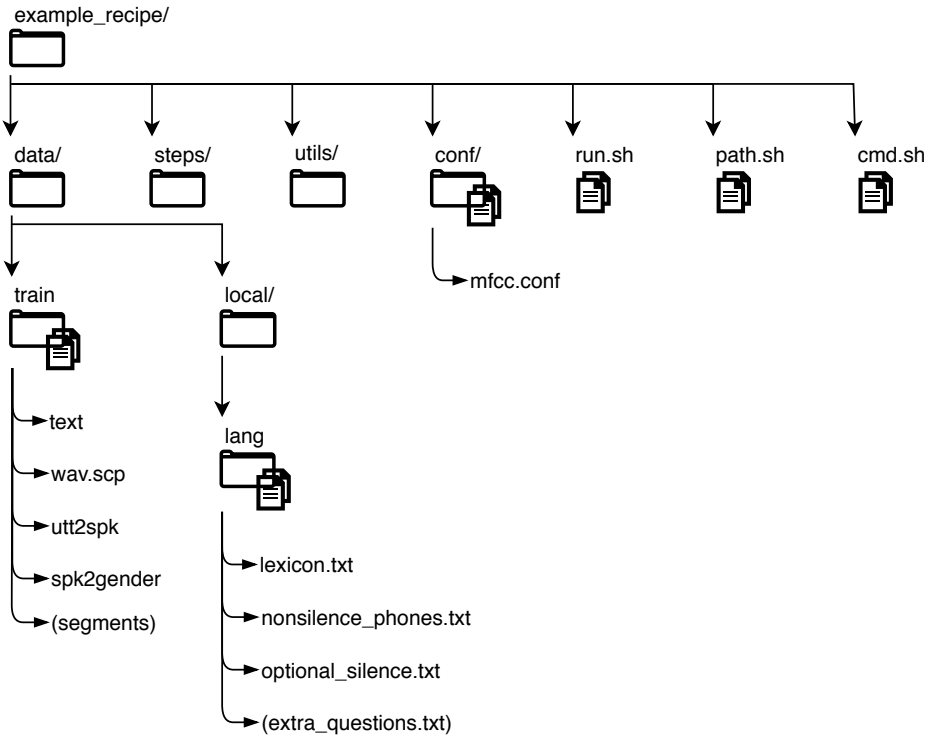


*Figure 3.8: Folder structure for starting Kaldi experiments.*

**Kaldi**
The top level of Kaldi consists of bash scripts, with which one can perform tasks for speech recognition systems. Existing recipes can be used to perform tasks in Kaldi or the scripts can be used for a speech recognition task with own data. The task of forced alignment consisted of different steps. First, the feature extraction was performed with a cepstral mean voice tract normalisation (CMVTN) afterwards. The language model was calculated with the help of the SRILM software (Stolcke, 2002). Although a language model is not necessary for the forced alignment task, all components were created in order to create full decoding graphs and during the training of the acoustic model the language model is reduced to a linear acceptor. All material from the CS component of the GRASS corpus was used to train a monophone acoustic model. After training the acoustic model, the data was aligned to the model and the resulting alignments were converted to `.ctm`, which was the starting point for the posprocessing part.

**Postprocessing**
Once the alignments were calculated with the *Kaldi* software. I analysed them with *Python*. I converted the Kaldi output to Praat using a similar method than Chodroff (2018). In contrast to the tutorial, I used a single programming language to make the process of converting the alignments to *Praat* more robust. The alignments were represented as *pandas dataframes* in order to achieve a user-friendly and more flexible format. Another advantage of *pandas dataframes* is the support of many vectorised processing techniques, which saves time for huge amounts of data. I created Python scripts to evaluate the alignments by comparing them with reference alignments, I calculated statistics about the lexicon and, as a last step, converted them into Praat TextGrids.

## 3.6 Evaluation of automatic speech segmentations

I evaluated the phonetic speech segmentations by comparing the automatic produced segmentations to reference segmentations. It is a common approach to analyse the results of an automatic speech segmentation task with the help of reference transcripts as described in (C. Van Bael et al., 2007). As reference alignments are done by human-annotators, there is always a subjective factor which leads to a certain variance of resulting phone symbols and phone boundaries. Goldman (2011) found the agreement between a machine alignment and each human alignment to be comparable to the inter-human agreement. Nevertheless, phonetic speech segmentations, done with the forced alignment task, is compared to reference segmentations with the help of the Levenshtein algorithm. The Levenshtein algorithm finds the minimum number of operations to convert a phone sequence into the reference phone sequence. Analysing the operations in a quantitative way allows a more detailed insight of the automatic phonetic speech segmentation process. Figure 3.9 shows an example of the Levenshtein algorithm by comparing an automatic speech segmentation with a reference segmentation of an utterance of the GRASS RS component.

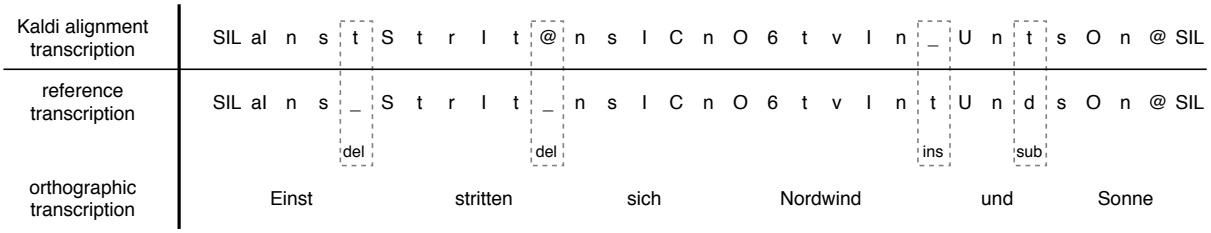


*Figure 3.9: Comparison of phone sequences with the Levenshtein distance.*

Note that SIL represents the phone for silence. The overall Levenshtein distance, also called disagreement, between two phone sequences is the sum over all operations related to the number of reference symbols:

$$Levenshtein\ distance = \frac{\#deletions + \#insertions + \#substitutions}{\#reference\ symbols} \cdot 100\% \tag{3.2}$$

A similar method of comparing phone sequences is the Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT) (Elffers, C. Van Bael, and Strik, 2005), which takes articulatory effects into account. Here I used the Levenshtein distance.

# 4 Forced Alignment of GRASS corpus

## 4.1 Validation of read speech alignments

For the RS component of the GRASS corpus, there exist reference alignments for a set of spoken utterances. In total there are 887 utterances from 38 different speaker for which reference alignments were done by human annotators. The correction by the annotators were done after an automatic speech segmentation of the utterances with the help of the WebMAUS tool (Kisler, Uwe Reichel, and Schiel, 2017). As described in Section 3.6, the difference between the automatic speech segmentations with the help of Kaldi and the manual aligned segmentations was calculated with the Levenshtein distance. Figure 4.1 shows a comparison of the Levenshtein distance between automatic speech segmentations that were done with different lexicons, and the reference alignments.

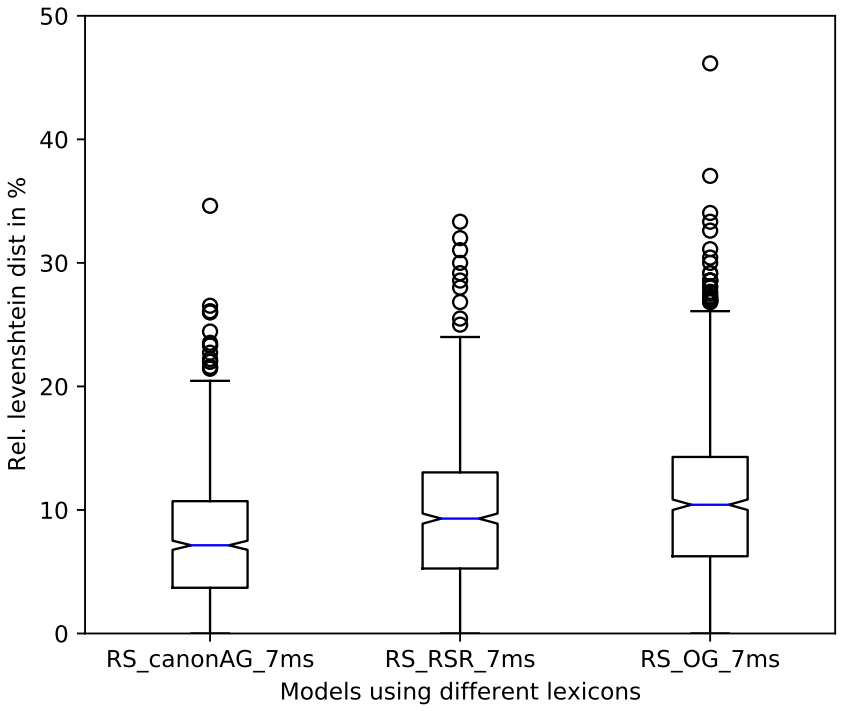


*Figure 4.1: Comparison between Kaldi alignments and reference alignments, when using different lexicons. Explanation of abbreviations: canonAG - canonical Austrian German, RSR - read speech rules, OG - overgeneralised.*

One can see that the FA model with the canonical Austrian German lexicon has a smaller distance to the reference alignments than FA models with a lexicon that contains pronunciation variants. Note that the starting point for the manually corrections for the reference alignments

is an automatic speech segmentation done with the WebMAUS tool, which is more similar to the FA model with the canonical Austrian German lexicon. A more detailed comparison is the analysis of the most frequent mismatches between the FA by Kaldi and the reference alignments in Table 4.1.

*Table 4.1: Most frequent mismatch between the FA models with a frame-shift of 7ms and reference alignments, when using different lexicons.*

| **phone symbol** | **relative count in %** | **op.** | **phone symbol** | **relative count in %** | **op.** | **phone symbol** | **relative count in %** | **op.** |
|---|---|---|---|---|---|---|---|---|
| @ | 30.7 | del | @ | 19.7 | del | t → d | 17.4 | sub |
| t | 10.2 | del | t → d | 18.5 | sub | @ | 9 | ins |
| r | 6.6 | del | p → b | 5.7 | sub | p → b | 7.9 | ins |
| d | 3.7 | del | SIL | 5 | ins | @ | 7.4 | del |
| 6 | 3.3 | del | 6 | 3.3 | del | SIL | 4.5 | ins |
| lexicon canonical Austrian German | | | lexicon read speech rules | | | lexicon overgeneralised | | |

Table 4.1 shows the most frequent operations, one has to perform to convert the Kaldi FA with different lexicons to the reference alignments. The most frequent mismatches for the Kaldi FA model with a canonical Austrian German lexicon are deletion operations, whereas the the most frequent mismatches for Kaldi FA models with pronunciation variants also contain substitution and insertion operations. Taking into account that a substitution of articulatory similar phones has a smaller distance than deletion or insertion operations, one can argue that the distance to the reference alignments is smaller with Kaldi FA models with pronunciation variants than with the canonical Austrian German lexicon.

Figure 4.2 illustrates the distance to the reference alignments for Kaldi FA models calculated with different frame-shifts. The Kaldi FA model, calculated with a frame-shift of 10 ms, has the smallest median of the Levenshtein distance. Table 4.2 summarises the most frequent Levenshtein operations for a FA with an overgeneralised lexicon and different frame-shifts.

*Table 4.2: Most frequent mismatches between the FA model with an overgeneralised lexicon and reference alignments, when using different frame-shifts.*

| **phone symbol** | **relative count in %** | **op.** | **phone symbol** | **relative count in %** | **op.** | **phone symbol** | **relative count in %** | **op.** |
|---|---|---|---|---|---|---|---|---|
| t → d | 18.5 | sub | t → d | 17.4 | sub | t → d | 15.8 | sub |
| @ | 8.6 | ins | @ | 9 | ins | @ | 8 | del |
| @ | 8.5 | del | p → b | 7.9 | sub | @ | 7.6 | ins |
| SIL | 6.6 | ins | @ | 7.4 | del | SIL | 6.1 | del |
| o → a | 5.8 | sub | SIL | 4.5 | ins | p → b | 4.8 | sub |
| frame-shift 10ms | | | frame-shift 7ms | | | frame-shift 5ms | | |

Taking into account that substitutions of articulatory similar phones have a smaller distance than deletion and insertion operations, a smaller distance between the reference alignments and the Kaldi FA can be reached with a frame-shift of 7ms.

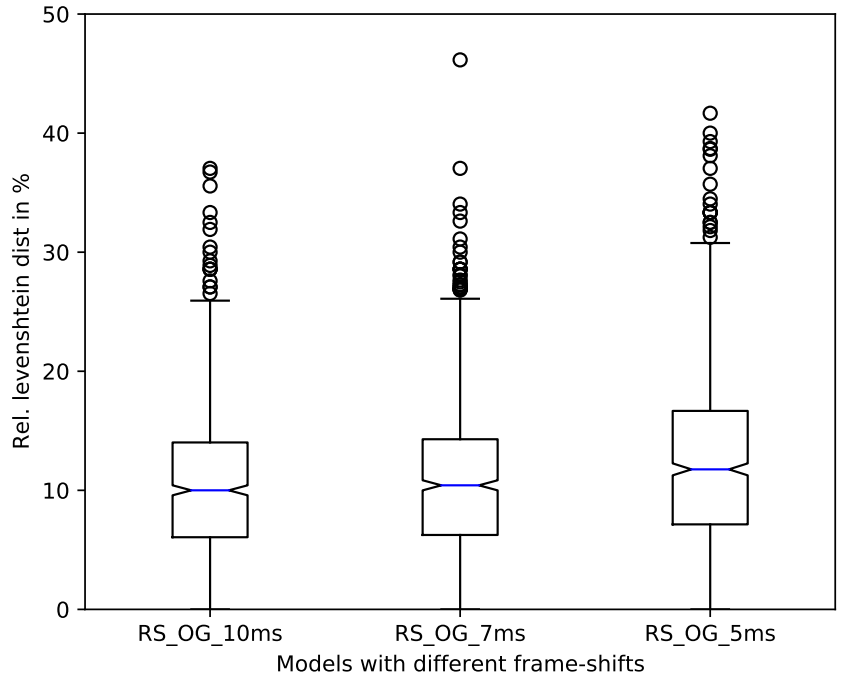


*Figure 4.2: Comparison between Kaldi FA models and reference alignments, when using different frame-shifts and an overgeneralised lexicon.*

## 4.2 Qualitative analysis of a read speech example

After the forced alignment task in Kaldi, I converted the alignments to the Praat TextGrid format. Figure 4.3 shows an example of an utterance from the RS component within the Praat software. There exist four different annotation layer, also called tiers. The first three tiers represent a phone segmentation for FA models that were calculated with different lexicons and the last tier shows the spoken sentence. The waveform and the spectrogram of the speech signal are plotted above the segmentation layers. In Figure 4.3, some phone boundaries of the different FA models vary among each other. The FA model with the overgeneralised lexicon and the read speech rules lexicon use a voiceless plosive /p/ instead of the voiced plosive /b/ for the FA of the word BLAUEN. Note that the phone boundary of the plosive /p/ in the second annotation layer is placed before the actual plosive sound happens, as you can see in the spectrogram. Also the phone start boundary of the first phone /a/ is placed into a silence section of the audio data in the second annotation layer. In contrast to the other alignments, the alignment with the overgeneralised lexicon annotates the word HIMMEL as /h I m l/ with a deletion of the schwa phone. The phone start boundaries of the phone /ts/ of the word ZIEHEN is placed into the sound of the following /l/ phone for the first two phone tiers. Only in the third phone tier, the phone start boundary of /ts/, is placed after the previous /l/ sound.

The next Figure 4.4 illustrates a Praat example of the same RS utterance with phone segmentations that are calculated with different frame-shifts and with an overgeneralised lexicon. An obvious error of the FA with a frame-shift of 10 ms and 5 ms is the first phone /o/ of the word AM, for which the phone boundaries were set inaccurately into silence parts. The vowel substitution from the canonical pronunciation /a m/ to /o m/ is also not equivalent to the reference alignment. The plosive burst of the phone /p/ of the word BLAUEN is correctly taken into account for the FA with a frame-shift of 10 ms and 7 ms, but the first one annotates the phone

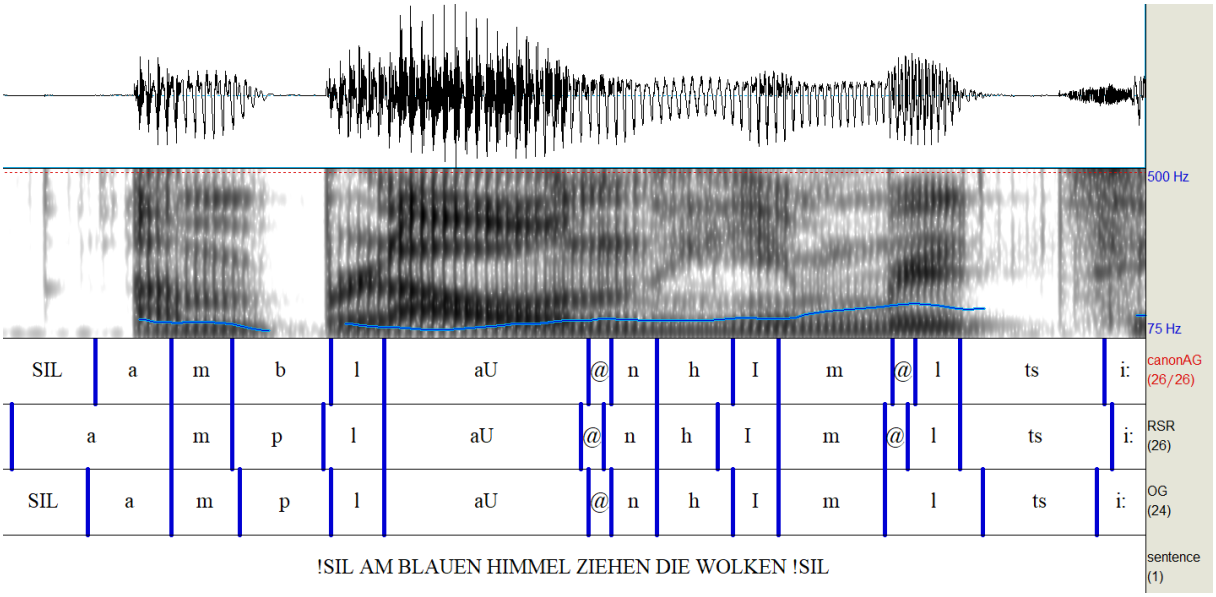


*Figure 4.3: Praat example of a RS utterance with Kaldi alignments, calculated with different lexicons and a frame-shift of 7ms.*

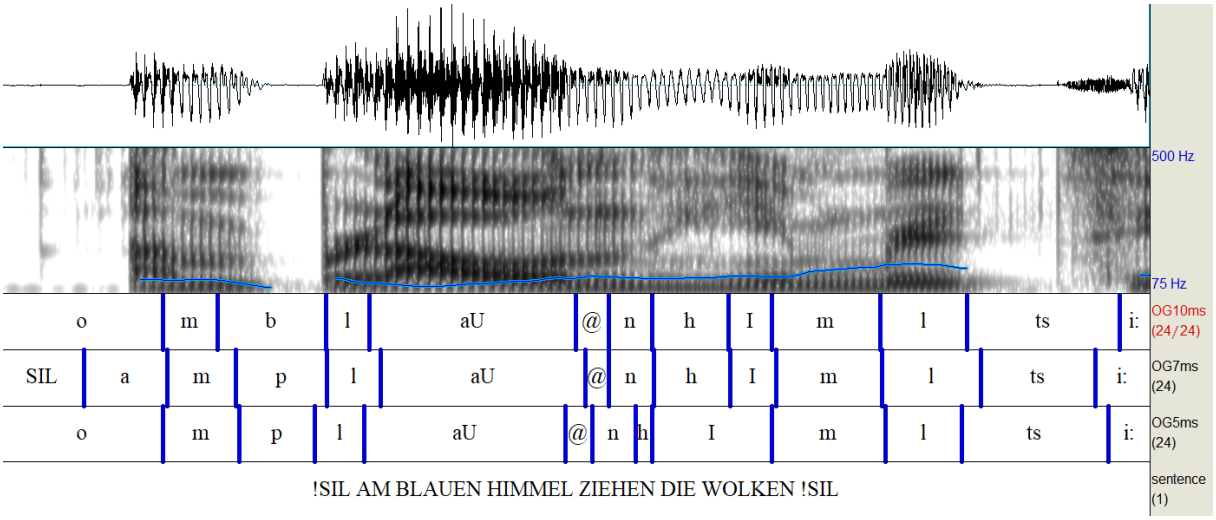


*Figure 4.4: Praat example of a RS utterance with Kaldi FA, calculated with different frame-shifts and an overgeneralised lexicon.*

as a voiced sound whereas the second alignment tier annotates the phone as a voiceless sound, which is more accurate. The boundaries of the phone /i/ of the word HIMMEL were not correctly set for the forced alignment model of the third tier and, as a consequence, the previous phone /h/ has an unreasonable short duration. All models detect a schwa deletion for the word HIMMEL, but the following sound /ts/ is best aligned for the alignment model with a frame-shift of 7 ms.

## 4.3 Validation of conversational speech alignments

There are no reference alignments for the CS component of the GRASS corpus, so an quantitative evaluation with a distance measure to reference data is not possible. Nevertheless, a distance measure was done between Kaldi alignments, which were calculated with different acoustical models. The first model was a monophone acoustical model, which is similar to the RS acoustical

models and the second model was a triphone based acoustical model. For more information about the used acoustical models see Section 3.4. Table 4.3 summarises the most frequent mismatches between the different FA models.

*Table 4.3: Most frequent mismatches between a FA with a monophone model and a FA with a triphone model, both with a frame-shift of 10 ms and a frame-length of 30 ms.*

| **phone symbol** | **relative count in %** | **op.** |
|---|---|---|
| SIL | 18.1 | del |
| SIL | 13.3 | ins |
| t | 4.7 | ins |
| o → a | 4.5 | sub |
| a → o | 4.4 | sub |
| @ | 3.2 | ins |
| d | 2.9 | ins |

It is interesting that the most frequent mismatches were silence deletions and silence insertions and leads to the presumption that the phone boundaries were different for the monophone alignment than for the triphone alignment, as different placed silence phones also influence the neighbouring phone boundaries. The substitutions between the articulatory similar phones /a/ and /o/ indicate that the decision of an /a/ or an /o/ sound is not consistent for different acoustical models. The next section shows a qualitative analysis of an example from the CS component for different acoustic models.

## 4.4 Qualitative analysis of a conversational speech example

Figure 4.5 illustrates a CS example in Praat for two phone alignments, which are calculated with different acoustical models.

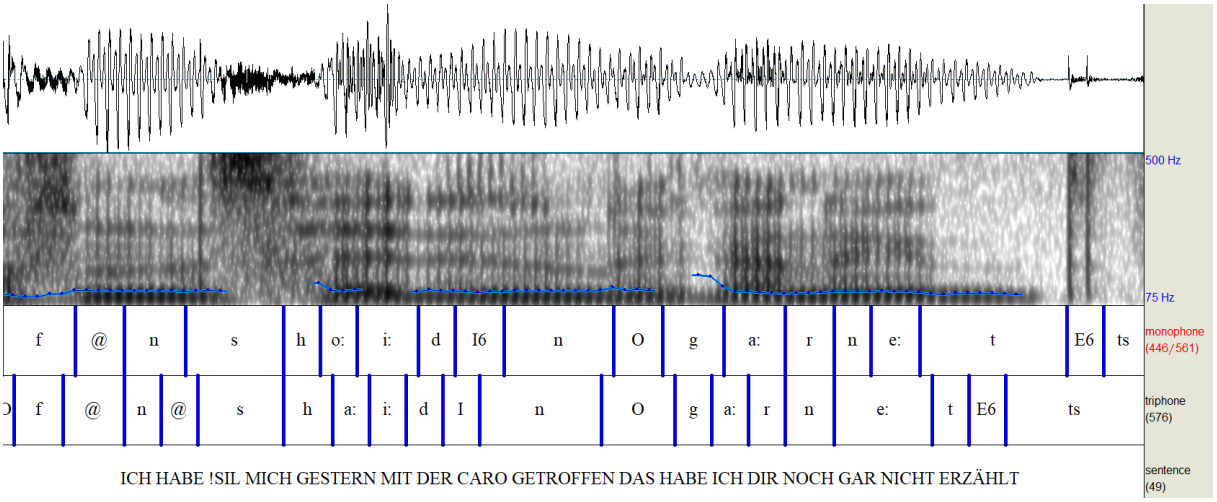


*Figure 4.5: Praat example of a CS utterance with a monophone FA model and a triphone FA model, both with a frame-shift of 10 ms and a frame-length of 30 ms.*

When comparing the two models, it is noticeable that almost all phone boundaries differ. The monophone FA model annotates the word DAS as a single phone /s/, which is a more reduced form

compared to the triphone FA model. As the utterance is spoken in a fast manner a reduced form is more consistent. The pronunciation variant for the word HABE differs for the two FA models. The triphone FA model annotates the word HABE as /h a:/, which is more consistent with the audio data. Another difference between the FA models is the segmentation of the words GAR NICHT. Although both models annotate the same pronunciation variants, the placement of the phones differ. The triphone FA model annotates the phone /t/ of the pronunciation /n e: t/ with a short duration, whereas the monophone FA model annotates a relatively long duration of the phone /t/. The segmentation of the word NICHT is more consistent to the audio data for the monophone FA model.

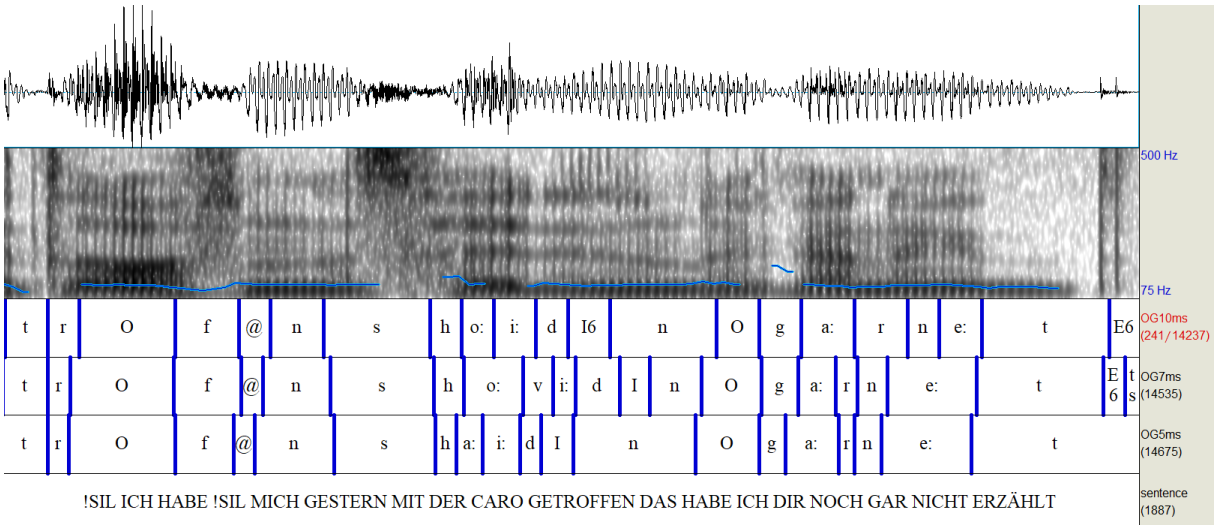


*Figure 4.6: Praat example of a CS utterance with a monophone alignment models with an overgeneralised lexicon, using different frame-shifts.*

Figure 4.6 illustrates a phone segmentation for different FA models of a CS utterance. The FA models were calculated with different frame shifts and an overgeneralised lexicon. All models annotate a different word pronunciation for the word HABE and as a result, also the phone boundaries of the following pronunciation /i:/ is placed differently for the FA models. While listening to the audio data, it is difficult to distinguish between the word boundaries for the word sequence HABE ICH DIR NOCH. Also the FA models annotate different phone sequences and phone boundaries in this section of the example utterance.

## 4.5 Estimation of pronunciation probabilities

After an automatic speech segmentation with an overgeneralised lexicon, I counted the number of word pronunciations used in the CS component. The overgeneralised lexicon contained 72 588 word pronunciations with canonical and varying pronunciations. After counting the occurrences of the word pronunciations, 18 342 word pronunciation entries occurred at least one time in the CS data. By comparing the word pronunciations and the lexicon, I calculated statistics about the pronunciation lexicon by counting occurring word pronunciations from the FA output. Table 4.4 shows an example of the lexicon statistic.

It shows that the canonical word pronunciation of the word GEBEN did not occur in the training data, whereas the most probable word pronunciation for the word GEBEN was the word pronunciation /g e: m/, which is the most reduced word pronunciation. It is interesting to investigate the usage of varying word pronunciations per speaker and to compare the results for different speaking styles.

*Table 4.4: Section of pronunciation lexicon with calculated statistics of GRASS CS component.*

| word | word pronunciation | absolute occurrence (count) | relative occurrence in % |
|---|---|---|---|
| ... | | | |
| ABER | a: b 6 | 1383 | 59.22 |
| ABER | o: b 6 | 952 | 40.77 |
| ... | | | |
| DAS | d e: s | 2893 | 44.37 |
| DAS | s | 1905 | 29.22 |
| DAS | @ s | 1370 | 21.01 |
| DAS | d a s | 351 | 5.38 |
| ... | | | |
| GEBEN | g e: m | 19 | 67.85 |
| GEBEN | g e: v @ n | 4 | 14.28 |
| GEBEN | g e: v n | 4 | 14.28 |
| GEBEN | g e: b m | 1 | 3.57 |
| GEBEN | g e: b @ n | 0 | 0.0 |
| GEBEN | g e: b n | 0 | 0.0 |
| ... | | | |

## 4.6 Comparison of different speaking styles

Starting point for creating the pronunciation variants are the canonical word pronunciations. With the help of the phonological rules, the canonical word pronunciations are changed and added as pronunciation variants to the pronunciation lexicon. By counting the used word pronunciations per speaker and categorise them into canonical word pronunciation or varying word pronunciation, it is possible to calculate the amount of varying word pronunciations per speaker. In Figure 4.7, the varying word pronunciation usage per speaker is compared for the RS component and the CS component. Both FA models used an overgeneralised + canonical AG lexicon and a frame-shift of 10 ms. As I used different mechanisms to create the pronunciation lexicons, the complexities of the overgeneralised lexicons slightly differ. The complexity of the RS overgeneralised lexicon was 3.4, whereas the complexity of the CS overgeneralised lexicon was 4.1. Note that the RS overgeneralised + canonical AG lexicon contained two rules less than the overgeneralised + canonical AG lexicon for CS, as the two rules were not applied once in the RS component. Except for speaker 029F, all speakers have a larger varying pronunciation usage for CS than the varying pronunciation usage for RS. In a slightly different representation, the impact of the different speaking styles on the usage of pronunciation variants can be seen even more obviously.

Figure 4.8 shows the varying pronunciation usage for different speaking styles per speaker in a box plot representation. It is clearly visible that the varying pronunciation usage is significantly higher for the CS component than the varying pronunciation usage for the RS component. The median value for the RS component is 32.78%, whereas the median value for the CS component is above fifty percent at 50.53%. Also the absolute variances for the varying pronunciation usage per speaker is higher for the CS component than for the RS component.

A different comparison between the speech styles is the speechrate, which is the number of phones per second. Figure 4.9 shows a normalised histogram for the speechrate of RS and CS speech style. The mean speechrate for CS speech with $\mu = 11.44$ $^{\text{phones}}/_{\text{second}}$ is more than twice as large as the mean speechrate for RS with $\mu = 4.3$ $^{\text{phones}}/_{\text{second}}$. Also the variances

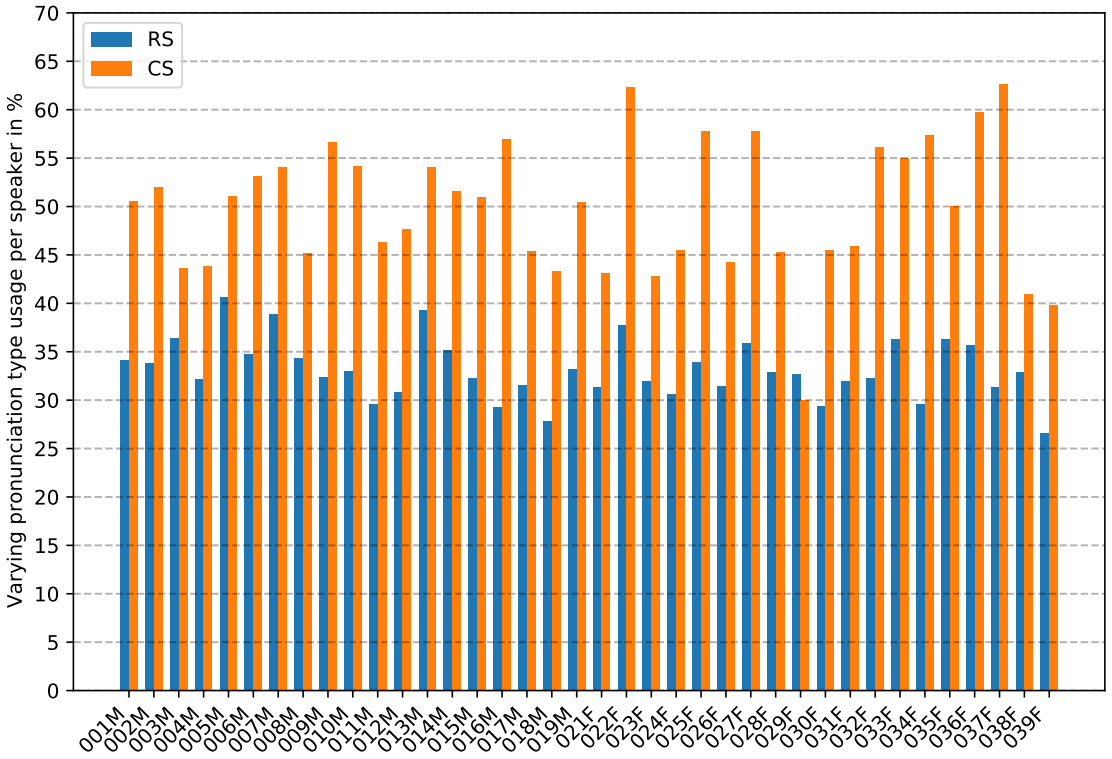


*Figure 4.7: Amount of pronunciation variant usage for different speaking styles, when using an overgeneralised + canonical Austrian German lexicon per speaker.*

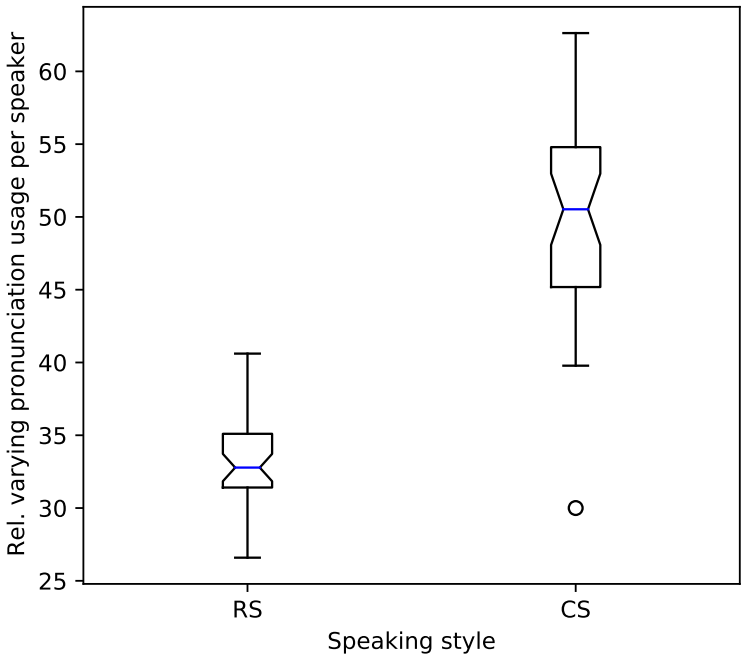


*Figure 4.8: Amount of pronunciation variant usage for different speaking styles, when using an overgeneralised lexicon.*

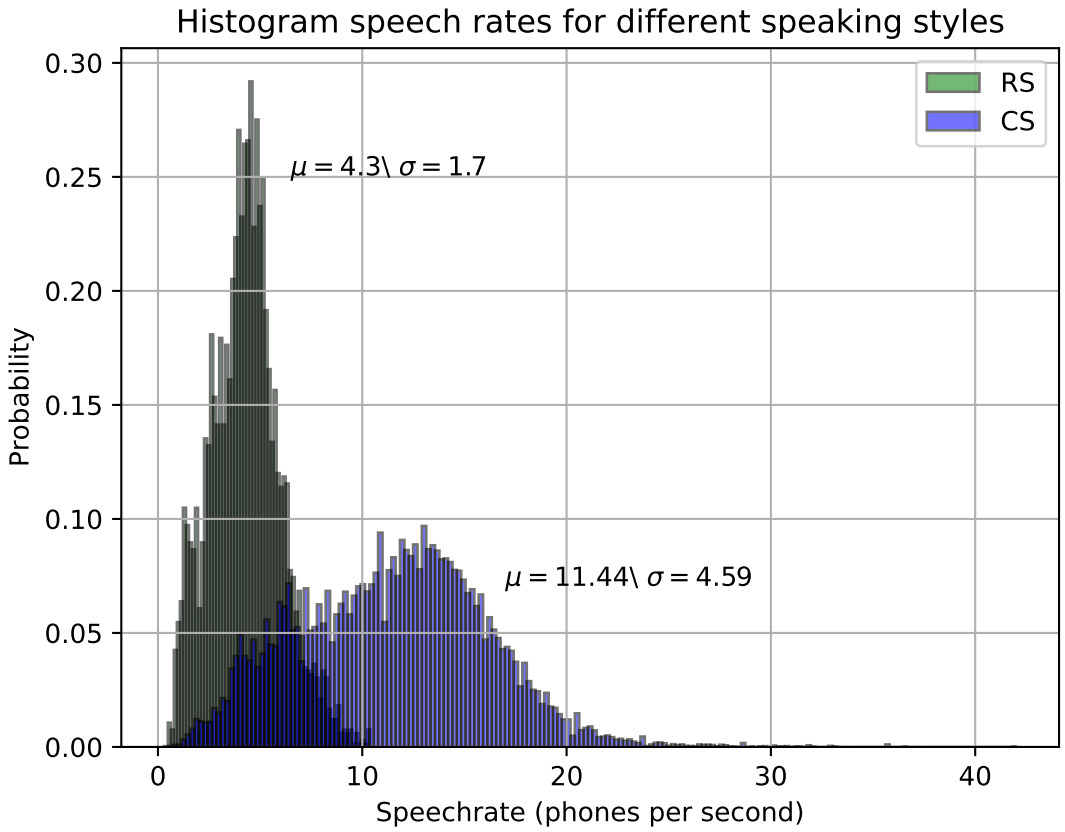


*Figure 4.9: Histogram of speechrates for different speaking styles.*

behave in a similar way. Next, I investigated speechrate relations per speaker. First, the mean speechrates and the variances of the speechrates were calculated for both speaking styles per speaker. Figure 4.10 illustrates the ratio between the mean speechrate for CS and the mean speechrate for RS per speaker. Speaker 018M has the lowest ratio between the mean speechrates for different speaking styles with a value of 2. This means that the average speechrate of speaker 018M is twice as large for CS compared to the speechrate of RS. Speaker 034F has the largest ratio between the mean speechrates with a value of 3.6. In summary one can say that for all speakers the mean speechrate for CS is more than twice as large than the mean speechrate for RS.

Figure 4.11 shows the ratio between the variance of the speechrate for CS and the variance of the speechrate for RS per speaker. As all ratios are greater than 1, one can state that for all speakers the variance speechrate for CS is more than four times larger than the variance speechrate for RS. Noticeable is the ratio of speaker 017M with a value of 13.1, which indicates that the variance of the speechrate for CS is very distinct to the variance of the speechrate for RS.

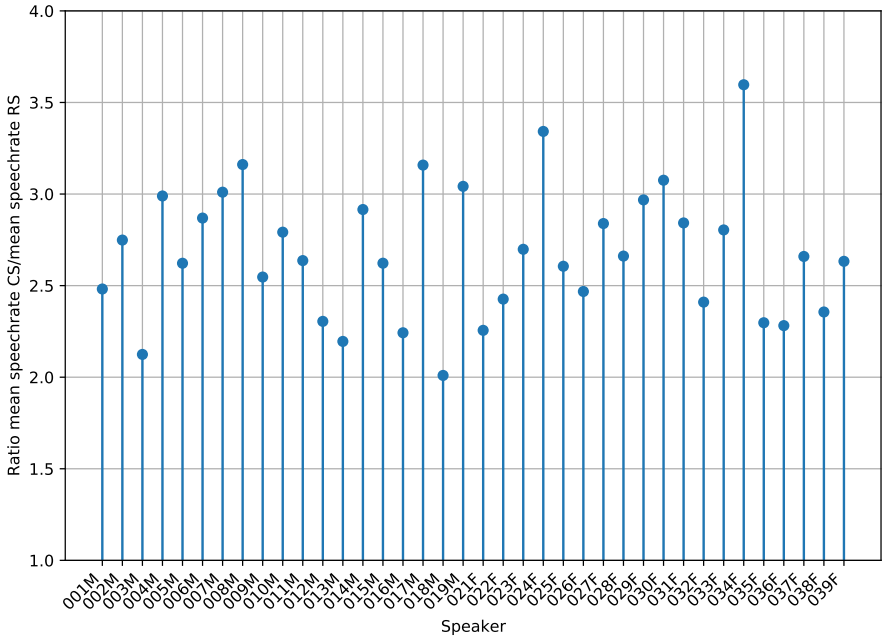


*Figure 4.10: Ratio between mean speechrate CS and mean speechrate RS per speaker.*

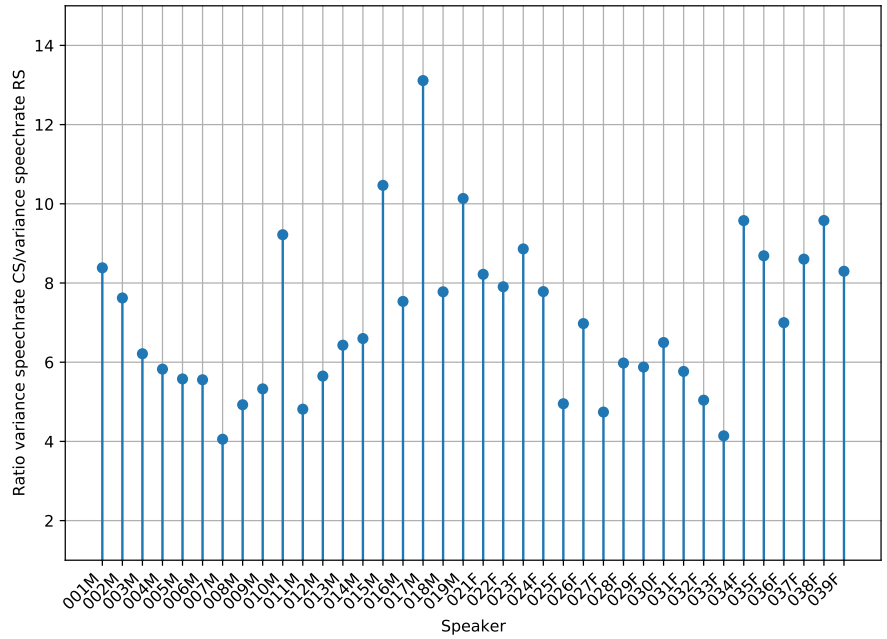


*Figure 4.11: Ratio between variance speechrate CS and variance speechrate RS per speaker.*

The comparison of the speech styles shows us that the pronunciation variant usage and the speechrate per speaker differ for RS compared to CS. The median pronunciation variant usage for RS with 32.78% is significantly smaller then the median pronunciation variant usage for CS with 50.53%. The variance of the varying pronunciation usage is also smaller for the RS component than for the CS component. Thus, it is more likely to find varying pronunciations during CS speaking style than for RS speaking style. Another indicator for the comparison is the speechrate measured as phones per second. The analysis per speaker showed that the overall speechrate for CS is more than twice as large as the overall speechrate for RS. Furui et al. (2005) also reported an increased speechrate for spontaneous speech compared to RS using the Corpus of Spontaneous Japanese (CSJ). The median pronunciation variant usage for CS with 50.53% is comparable to the reported value for the citation word pronunciation of 56% for Dutch language (Schuppler, Ernestus, et al., 2011). The comparison of the pronunciation variant usage and the speechrates between RS and CS showed that there is a larger amount of variability in CS than in RS. Finke and Waibel (1997) stated a larger amount of variability due to accents, speaking style and speechrates in the Switchboard corpus. The authors proposed a speaking mode dependent pronunciation modelling in order to improve recognition results.

# 5 Discussion & Conclusion

The aim of this thesis was to perform an automatic speech segmentation of the GRASS corpus. For the automatic segmentation of the corpus, it is relevant to train an acoustic model with the corpus data. Although the MAUS tool offers different language settings for German, there is no language setting for Austrian German. The MFA and the SPPAS tool offer the possibility to use a new language or to train the acoustic model with new data. Nevertheless, it is possible to use the MAUS tool with the German language setting to perform an automatic speech segmentation task for the GRASS corpus. For the RS component, the automatic segmentation from the MAUS tool is quite satisfying and serves as starting point for reference alignments. However, the CS component of the GRASS corpus is more challenging due to the more complex speaking style and there is a need to model the pronunciation variations of Austrian German in an own forced alignment system. Bigi and Meunier (2018) address the challenging task of within speech during spontaneous speaking style and incorporate knowledge about events of filled pause, laughter and noise into the acoustic model of the SPPAS tool for the French language. The work of Schiel (2015) explains that the statistical model consists of an acoustical model and an additional apriori probability model. He argues that automatic segmentation and labelling systems differ in terms of search space and in the way of how the apriori pronunciations are modelled. In summary, the existing forced alignment systems have different advantages and it would be of interest to combine the mechanisms that are relevant for CS.

Strik and Cucchiarini (1999) presented a decision based framework in order to characterise the pronunciation modelling in an ASR system. As the task of automatic speech segmentation is closely related to the ASR task, I used this decision based framework to characterise the pronunciation modelling process of Austrian German. In this thesis the pronunciation variation of Austrian German in a CS style was modelled. The intraspeaker variation within spontaneous speech contained assimilation, reduction, deletion and insertion phenomena. With the help of a pronunciation lexicon, the word-internal pronunciation variation was modelled. The information of the pronunciation variation of Austrian German came from the study of Schuppler, Adda-Decker, and Morales-Cordovilla (2014) and is a knowledge-based approach. In contrast to the knowledge-based approach, a data-driven approach was not possible for the CS component of the GRASS corpus as the data is not segmented on a phone or word level. With the usage of phonological rules, the information of the pronunciation variation was represented in a formalised way. The advantage of formalised information is that a quantitative analysis of the rules used gives insight to the pronunciation variation of the training data. Such a study has been done by Schuppler, Adda-Decker, and Morales-Cordovilla (2014) for parts of the GRASS corpus with the help of HTK experiments. The pronunciation variation is modelled in the component of the lexicon with a pronunciation lexicon with variants in an explicit way. In contrast to an ASR task, for a forced recognition task it is not useful to model the pronunciation variation in an implicit way within the acoustic model, as one is interested in the direct result of the pronunciation variation, which is the best fitting phone sequence to given audio data.

I analysed the automatic speech segmentations of the GRASS corpus for the RS component and the CS component in a quantitative and a qualitative way. For a part of the RS corpus

data, there are reference alignments available, with which I compared the resulting automatic speech segmentations. The distance measure between the calculated segmentations and the reference alignments was done with the Levenshtein algorithm. The comparison of different models, calculated with different lexicons, showed that the smallest Levenshtein distance can be achieved with a canonical Austrian German model and a median value of 8.4%. C. P. J. Van Bael (2007) reported a mean value of a distance measure of a similar segmentation approach by forced recognition to reference alignments with a value of 8.1%. Note that the distance measure in the study of C. P. J. Van Bael (2007) was calculated with the ADAPT algorithm, which differs to the Levenshtein distance by taking articulatory effects into account. By representing the most frequent mismatches between the different models and the reference alignments the results showed that some of the most frequent mismatches are substitutions between articulatory similar phones. Anticipating that substitutions of articulatory similar phones would have a smaller distance, by using an algorithm like ADAPT, the best alignment model for RS would be a model with a lexicon with read speech rules. This is a lexicon which is calculated with a subset of the phonological rules and has a smaller complexity than an overgeneralised lexicon. In the qualitative analysis of a RS example the model with the overgeneralised lexicon showed the most promising alignment results. Furthermore, models with different feature extraction mechanisms were compared. The acoustic models were trained with a feature extraction, where different frame-shifts are used. The quantitative analysis showed that the differences between the models, compared to the reference models, were quite small and the lowest median value was calculated with a frame-shift of 10ms with a Levensthein distance of 8.5%. In a qualitative analysis between models with different frame-shifts, a model with a frame-shift of 7ms is the most promising one.

For the CS component of the GRASS corpus, a comparison between a monophone based forced alignment model and a triphone based forced alignment model was calculated with the help of the Levenshtein distance. With an overall Levensthein distance of 12.6%, there is a significant difference between a monophone based forced alignment model and a triphone based alignment model. The most frequent mismatches include deletion and insertion of silence phones as well as substitutions between the phones /a/ and /o/. With different placements of silence phones, also neighbouring phones were effected and I expected a large difference in the phone boundaries, when comparing the monophone based model and the triphone based alignment model. The qualitative analysisis of a CS example showed that there is a significant difference both with respect to phone symbols and phone boundaries.

An analysis after performing the automatic segmentation task with an overgeneralised lexicon is to estimate the pronunciation probabilities. For the CS component, the results showed that from the original overgeneralised lexicon with 72 588 word pronunciation entries, only 18 342 word pronunciations are actually used in the corpus data. By grouping the pronunciation lexicons into canonical word pronunciations and varying word pronunciations it was possible to analyse the usage of varying word pronunciations compared to canonical word pronunciations per speaker.

It is of interest to compare the RS speech style to the CS speech style in order to gain a deeper understanding of spontaneous speech. In section 4.6, the usage of varying word pronunciations per speaker was calculated for the RS component and the CS component. The median for the varying word pronunciation usage per speaker for the RS component is 32.8% and is significantly lower than the median of the varying word pronunciation usage per speaker for the CS component with 50.5%. A similar study for the Dutch language reported an overall value for the citation word pronunciation of 56% (Schuppler, Ernestus, et al., 2011). A second analysis compared the speechrates of the CS and RS components of the corpus. The overall mean value of the speechrate for RS was 4.3 $^{\text{phones}}/_{\text{second}}$, whereas the overall mean value of the speechrate

for CS is 11.4 $^{\text{phones}}/_{\text{second}}$. Also the variances of the speechrate for the different speaking styles showed similar behaviour. An analysis of individual speakers showed that for all speakers the speechrate for CS is more than twice as high as the speechrate for RS. The biggest difference between the speechrate for CS compared to RS, measured for one speaker, was a factor of 3.6.

Summing up, the automatic speech segmentation of the GRASS corpus consisted of different challenges. The quantitative evaluation of the RS component showed that the usage of a lexicon with pronunciation variants yielded similar results than a lexicon without pronunciation variants. The FA for the CS component was calculated with an overgeneralised pronunciation lexicon in order to cover the pronunciation variability of the CS. The automatic speech segmentation was a starting point for the analysis of the speechrate and the pronunciation variation usage. A comparison between RS and CS in regard to speechrates and pronunciation variation usage showed that there is a larger amount of variability in CS than in RS. As a result the usage of a pronunciation variant lexicon in a FA for CS is important to cover the large amount of variability.

# 6 Outlook

One focus of this thesis was the modelling of the pronunciation variants of Austrian German. In the process of creating such a pronunciation lexicon for the alignment of CS, the detailed information of the annotations from the GRASS corpus was used. The resulting lexicons were processed in different ways and merged afterwards. In addition to a foreign-language lexicon or a broken word lexicon it would be of interest to create a multi-word lexicon, containing the pronunciations of the annotated multi-words. With help of such a multi-word lexicon cross-word processes during CS could be also modelled.

Another focus of this work was the observation of using different frame-shifts during the feature extraction process. As Kaldi offers the possibility to use many triphone based acoustic models an evaluation of the automatic segmented alignments by using different triphone acoustic models would be interesting. The authors McAuliffe et al. (2017) report that using triphone based models and speaker adaptive training leads to a better automatic segmentation when comparing phone boundaries.

Similar to the quantitative evaluation of the RS component in section 4.1, an evaluation for the CS component would be of great interest. Starting point for a quantitative distance measure is the creation of reference alignments. Although the creation of human-labelled alignments is time-consuming, it would significantly improve the quantitative evaluation of the CS data. Another benefit during the evaluation process would be the usage of an algorithm that takes articulatory effects into account. C. Van Bael et al. (2007) developed the ADAPT algorithm to determine the distance between the automatic segmented alignments and the reference alignments. The advantage of using such an algorithm is that the analysis of the most frequent mismatches would list more relevant phone operations.

Section 4.5 describes the estimation of pronunciation probabilities for CS of the GRASS corpus. Motivated by improving an ASR task, the authors Chen et al. (2015) incorporate the knowledge of pronunciation probabilities into the lexicon. In addition to the pronunciation probabilities, they also investigate the impact of inter-word silence modelling. It would be interesting to analyse if the inter-word silence modelling and the incorporation of pronunciation probabilities have a positive effect on automatic speech segmentation.

# Bibliography


[AL99] Martine Adda-Decker and Lori Lamel. "Pronunciation variants across system configuration, language and speaking style." In: *Speech Communication* 29.2-4 (1999), pp. 83–98.

[Big12] Brigitte Bigi. "SPPAS: a tool for the phonetic segmentations of Speech". 2012.

[BM18] Brigitte Bigi and Christine Meunier. "Automatic segmentation of spontaneous speech". 2018.

[BW18] Paul Boersma and David Weenink. *Praat: doing phonetics by computer*. 2018. URL: http://www.%20praat.%20org/.

[Che+15] Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur. "Pronunciation and silence probability modeling for ASR." In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[Cho18] Eleanor Chodroff. *Corpus Phonetics Tutorial*. 2018. arXiv: 1811.05553. URL: http://arxiv.org/abs/1811.05553.

[Dav13] Namrata Dave. "Feature extraction methods LPC, PLP and MFCC in speech recognition." In: *International journal for advance research in engineering and technology* 1.6 (2013), pp. 1–4.

[DWP96] Neeraj Deshmukh, Mary Weber, and Joe Picone. "Automated generation of N-best pronunciations of proper nouns." In: *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 1. 1996, pp. 283–286.

[EVS05] Bram Elffers, Christophe Van Bael, and Helmer Strik. *Algorithm for Dynamic Alignment of Phonetic Transcriptions*. Tech. rep. Citeseer, 2005.

[Fur+05] Sadaoki Furui, Masanobu Nakamura, Tomohisa Ichiba, and Koji Iwano. "Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese." In: *Speech Communication* 47.1-2 (2005), pp. 208–219.

[FW97] Michael Finke and Alex Waibel. "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition." In: *Fifth European Conference on Speech Communication and Technology*. 1997.

[Gol11] Jean-Philippe Goldman. "EasyAlign: an automatic phonetic alignment tool under Praat." In: *Interspeech'11, 12th Annual Conference of the International Speech Communication Association*. 2011.

[GY08] Mark Gales and Steve Young. *The application of hidden Markov models in speech recognition*. 2008.

[Hai02] Thomas Hain. "Implicit pronunciation modelling in ASR." In: *ISCA tutorial and research workshop (ITRW) on pronunciation modeling and lexicon adaptation for spoken language technology*. 2002.

[Häm+09] Annika Hämäläinen, Michele Gubian, Louis ten Bosch, and Lou Boves. "Analysis of acoustic reduction using spectral similarity measures." In: *The Journal of the Acoustical Society of America* 126.6 (2009), pp. 3227–3235.

[Han+14] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. "Deep speech: Scaling up end-to-end speech recognition." In: *arXiv preprint arXiv:1412.5567* (2014).

[JDB18] Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. "Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data". 2018.

[KRS17] Thomas Kisler, Uwe Reichel, and Florian Schiel. "Multilingual processing of speech via web services." In: *Computer Speech & Language* 45 (2017), pp. 326–347. DOI: http://dx.doi.org/10.1016/j.csl.2017.01.005.

[Lam+03] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. "The CMU SPHINX-4 speech recognition system." In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*. Vol. 1. 2003, pp. 2–5.

[LKS01] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. "Julius—an open source real-time large vocabulary recognition engine". 2001.

[McA+17] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech*. Vol. 2017. 2017, pp. 498–502.

[MPR08] Mehryar Mohri, Fernando Pereira, and Michael Riley. "Speech recognition with weighted finite-state transducers." In: *Springer Handbook of Speech Processing*. Springer, 2008, pp. 559–584.

[PK17] Beat Pfister and Tobias Kaufmann. *Sprachverarbeitung*. 2017. DOI: 10.1007/978-3-662-52838-9.

[Pov+11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. "The Kaldi speech recognition toolkit." In: *IEEE 2011 workshop on automatic speech recognition and understanding*. CONF. IEEE Signal Processing Society. 2011.

[Rei12] U.D. Reichel. "PermA and Balloon: Tools for string alignment and text processing." In: *Proc. Interspeech*. Portland, Oregon, 2012, 4 pages.

[Ryb+11] David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltan Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. "Rasr-the rwth aachen university open source speech recognition toolkit." In: *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*. 2011.

[SAM14] Barbara Schuppler, Martine Adda-Decker, and Juan A Morales-Cordovilla. "Pronunciation variation in read and conversational Austrian German." In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

[SC99] Helmer Strik and Catia Cucchiarini. "Modeling pronunciation variation for ASR: A survey of the literature." In: *Speech Communication* 29.2-4 (1999), pp. 225–246.

[Sch+11] Barbara Schuppler, Mirjam Ernestus, Odette Scharenborg, and Lou Boves. "Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions." In: *Journal of Phonetics* 39.1 (2011), pp. 96–109.

[Sch15] Florian Schiel. "A statistical model for predicting pronunciation." In: *ICPhS*. 2015.

[Sch99] Florian Schiel. "Automatic phonetic transcription of non-prompted speech". 1999.

[SHZ17] Barbara Schuppler, Martin Hagmüller, and Alexander Zahrer. "A corpus of read and conversational Austrian German." In: *Speech Communication* (2017). ISSN: 01676393. DOI: 10.1016/j.specom.2017.09.003.

[Sto02] Andreas Stolcke. "SRILM-an extensible language modeling toolkit." In: *Seventh international conference on spoken language processing*. 2002.

[Van+07] Christophe Van Bael, Lou Boves, Henk Van Den Heuvel, and Helmer Strik. "Automatic phonetic transcription of large speech corpora." In: *Computer Speech & Language* 21.4 (2007), pp. 652–668.

[Van07] Christophe Patrick Jan Van Bael. *Validation, automatic generation and use of broad phonetic transcriptions*. UB Nijmegen [Host], 2007.

[VK11] Damjan Vlaj and Zdravko Kačič. "The influence of Lombard effect on speech recognition." In: *Speech technologies* 2014 (2011), pp. 1998–2001.

[YY93] Steve J Young and Sj Young. "The HTK hidden Markov model toolkit: Design and philosophy". 1993.

# A
# List of abbreviations

| | |
|---|---|
| **ADAPT** | Algorithm for Dynamic Alignment of Phonetic Transcriptions |
| **AG** | Austrian German |
| **ASR** | Automatic Speech Recognition |
| **CMVTN** | Ceptral Mean Voice Tract Normalisation |
| **CS** | Conversational Speech |
| **DCT** | Discrete Cosine Transform |
| **FA** | Forced Alignment |
| **FFT** | Fast Fourier Transform |
| **FST** | Finite State Transducer |
| **GMM** | Gaussian Mixture Model |
| **HMM** | Hidden Markov Model |
| **MFCC** | Mel Frequency Cepstral Coefficients |
| **RS** | Read Speech |
| **SGMM** | Subspace Gaussian Mixture Model |
| **WFST** | Weighted Finite State Transducer |

# B
# List of phones

| symbol | example word | transcription |
|---|---|---|
| *Plosive phones* | | |
| p | PAUKE | p aU k @ |
| b | BÜCHER | b y: C 6 |
| t | TÄNZE | t E n ts @ |
| d | DICKICHT | d I k I C t |
| k | KONSERVEN | k O n z E6 v @ n |
| g | GERÄUSCH | g @ r OY S |
| *Affricate phones* | | |
| ts | ZUCKER | ts U k 6 |
| dZ * | DSCHUNGEL | dZ U N @ l |
| tS | CHARLES | tS a: l s |
| *Fricative phones* | | |
| f | FRÜHJAHR | f r y: j a: r |
| v | WALZER | v a l ts 6 |
| s | WERKSFERIEN | v E6 k s f e:6 j @ n |
| z * | ANGESAGT | a n g @ z a: k t |
| S | SPIELE | S p i: l @ |
| Z * | GENIE | Z E n i: |
| C | LEICHT | l aI C t |
| j | NEUJAHR | n OY j a: r |
| x | OBACHT | o: b a x t |
| h | SCHÖNHEIT | S 2: n h aI t |
| *Sonorants* | | |
| m | MILCH | m I l C |
| n | NORDWIND | n O6 t v I n t |
| N | DING | d I N |
| l | LESEN | l e: z @ n |
| r | RUFEN | r u: f @ n |
| *Checked vowels* | | |
| I | SCHILD | S I l t |
| i | VIELLEICHT | f i l aI C t |
| E | WENN | v E n |
| e | RESIGNIERT | r e z I g n i:6 t |
| a | SACHEN | z a x @ n |
| O | SCHLOSSBERG | S l O s b 6 k |
| o | ABDREHEN | o p d r e: @ n |
| U | HUNDE | h U n d @ |
| u | AKTUELL | a k t u E l |
| Y | ZURÜCK | ts u r Y k |

| 9 | BÖLLERSCHUSS | b 9 l 6 S U s |
|---|---|---|
| **symbol** | **example word** | **transcription** |
| *Free vowels* | | |
| i: | FLIEGEN | f l i: g @ n |
| e: | FERIEN | f e:6 j @ n |
| E: | DÄNEMARK | d E: n @ m a r k |
| a: | KARTEN | k a: t @ n |
| o: | ROSA | r o: z a |
| u: | SCHUH | S u: |
| y: | SPÜLEN | S p y: l @ n |
| y | BÜROMÖBEL | b y r o: m 2: b @ l |
| 2: | ÖL | 2: l |
| *Free diphtongs* | | |
| aI | DREIRAD | d r aI r a: t |
| aU | LAUFEN | l aU f @ n |
| OY | HEUTE | h OY t @ |
| *Schwa vowel* | | |
| @ | GEWINNT | g @ v I n t |
| *Diphtongs* | | |
| 6 | SCHALTER | S a l t 6 |
| i:6 | VIER | f i:6 |
| I6 | VIERZEHN | f I6 ts e: n |
| y:6 | TÜR | t y:6 |
| Y6 | BÜRGER | b Y6 g 6 |
| e:6 | MEHR | m e:6 |
| E6 | LERNEN | l E6 n @ n |
| E:6 | NÄHER | n E:6 |
| 2:6 | STÖRT | S t 2:6 t |
| 96 | DÖRFCHEN | d 96 f C @ n |
| a:6 | FAHRKARTEN | f a:6 k a6 t @ n |
| u:6 | NUR | n u:6 |
| U6 | WURDE | v U6 d @ |
| o:6 | DOKTOR | d O k t o:6 |
| O6 | MORGEN | m O6 g @ n |

*Table B.1: List of phones with example words and pronunciations. Phones marked with * are omitted when the rule of silibant devoicing of the alveolar fricative /z/ is applied as a replacement rule.*

# C

# Phone mapping and G2P settings

| **phone set foreign language words** | eI | { | A: | e@ | O: | @U | Q | OI | V | w | R | 3: | ll | ttS | I@ | D | H | o∼ | rr | G | pp | a∼ | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **phone set Austrian German** | e I | E: | a:6 | e:6 | o:6 | O U | O | OY | a: | v | r | 2: | l | tS | i: @ | s | u | O | r | g | p | o: | s |

*Table C.1: Phone mapping from foreign languages to Austrian German phone set.*

| **transcription tag** | **G2P language setting** |
|---|---|
| <*ENGL> | eng |
| <*L> | deu |
| <*SP> | spa-ES |
| <*F> | fra-FR |
| <*SV> | swe-SE |
| <*I> | ita-IT |
| <*JAP> | spa-ES |
| <*HR> | deu |
| <*PT> | spa-ES |

*Table C.2: Mapping between transcription tags and G2P language settings.*