UNIVERSITY OF BONN

BACHELOR'S THESIS
BACHELORARBEIT

Data-Driven and Rule-Based Approaches to Improving Bavarian Speech Recognition

Tobias Bystrich
tobias_b@mailbox.org

First Examiner	*Second Examiner*	*Mentor*
Prof. Dr. Christian Bauckhage	Dr. Christoph Schmidt	Julia Pritzen, M.Sc.

Institut für Informatik,
Rheinische Friedrich-Wilhelms-Universität Bonn

Fraunhofer-Institut IAIS

2022-11-30
Corrected and extended version, 2023-03-23

Fraunhofer IAIS

UNIVERSITÄT BONN

Abstract

In this work we explored different approaches to improving the automatic speech recognition for the German dialect Bavarian for which little transcribed training data exists. The aim was to transcribe Bavarian dialect to written Standard High German. From the annotated data we could gather, we created datasets for training and testing. This resulted in approximately 24 hours of annotated training data, while the three test sets amount to approximately 6.7 hours in total. Our Bavarian models were based on the Fraunhofer IAIS hybrid model and two end-to-end models based on wav2vec 2.0. The experiments in this work may be classified using two categories: Rule-based approaches and data-driven approaches. While the wav2vec 2.0 models were trained in a purely data-driven manner, both approaches were applied to the hybrid model. The data-driven approach for the hybrid model also included training on approximately 24 hours of Bavarian speech. Additionally, a language model was trained on a larger set of transcriptions and subtitle data: a text corpus of annotations from the annotated training set and subtitles of a Bavarian television series. The rule-based approach was conducted via lexicon modifications using phonetic mapping rules. We found promising results and a relative -30% difference of WER compared to the hybrid baseline model. We find that despite Bavarian being a low-resource language, the greatest potential for improving Bavarian speech recognition lies in data-driven training of the acoustic modeling even though all approaches show promise.

Acknowledgements

This work was only possible due to the support of others. Therefore, I want to dedicate this section to these people. Since these personal words of gratitude would otherwise sound official and artificial, I hope that you'll understand that I will not use academic titles.

This thesis for the University of Bonn was written at and using the technologies of the Fraunhofer Institute IAIS, more specifically the NetMedia department.

I would like to deeply thank Julia Pritzen. Her friendly response to my interest in her published master's thesis was essential for me getting the chance to write this thesis now. As my mentor, her support enabled me to write the thesis the way I envisioned it. Apart from regular support and interest in my project she made it possible to benchmark the rule-based approaches locally without going through a complex institute-wide publishing process and helped make the Bavarian wav2vec 2.0 training a reality.

Secondly, I want to express my gratitude to Christoph Schmidt, Michael Gref and Josef Bauer who regularly followed my progress with interest: Christoph Schmidt for being my examiner, his support and knowledge and the opportunity to work at NetMedia; Michael Gref for his support which allowed me to get this opportunity and especially this topic for my thesis, as well as his technical expertise and knowledge about scientific work and ASR; and Josef Bauer for his time and interest in this work as much as for his knowledge and important advice.

Sincere thanks go to Christian Bauckhage for agreeing to be my examiner and thereby enabling this thesis at the University of Bonn.

I also would like to thank Bayerischer Rundfunk for their Bavarian broadcast data and Jonas Schreiber for the technical replies and efforts to make more data available, as well as the Forschungszentrum DSA for the linguistic interview data they supplied. Without this speech data, this work would have been impossible.

My colleagues and the IAIS team were most welcoming and open to my questions. While I cannot name all, I want to explicitly thank Ankita Oswal for supervising the beginning of my work and for her time and expertise regarding language models. Secondly, Tilo Himmelsbach without whom the wav2vec 2.0 training would not have been possible.

Thank you to Jan Seifert and Claudia Wich-Reif at the University of Bonn's German linguistics program. Jan Seifert patiently replied to my numerous mails and thereby helped to solve university-related issues that arose from my plan to work at Fraunhofer IAIS. Without this support, the thesis may not have been possible. Claudia Wich-Reif reviewed an early report on my work in the context of linguistics, provided an assessment of publicly available dialectal resources and gave many captivating university lectures on dialectology.

And last but definitely not least, I want to thank my good friend Lín, 赵琳, deeply for her support spending so much time reading my drafts and commenting on the structure of my text when all I could see were unfinished segments.

Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iii

1 Introduction 1

1.1 Bavarian Dialect 2

1.2 Problem Definition 2

1.3 Relevance 3

1.4 Objective and Research Questions 4

Research Questions 4

1.5 Limitations 5

2 Fundamentals 5

2.1 Linguistics 6

2.2 Phonetic Alphabets 8

2.3 Features of Bavarian 10

2.4 ASR Fundamentals 11

2.5 Low-Resource and Dialectal ASR 16

2.6 Methodology 17

2.6.1 Standard German Baseline Models 17

2.6.2 Benchmarking and Metrics 18

2.6.3 List of Approaches 19

3 Bavarian Data 20

3.1 Bavarian Data Collection 20

3.2 Creation of Datasets 21

3.3 Overview of Bavarian Datasets 23

4 Experiments 23

4.1 Rule-Based Lexicon Additions 24

4.2 Data-Driven Approaches 34

4.2.1 Training of a Hybrid Model 34
4.2.1a Acoustic Model 34
4.2.1b Language Model 35
4.2.1c Benchmarking 38
4.2.2 End-to-end Fine-Tuning using Wav2vec 2.0 41
5 Conclusion 47
6 Outlook 50
Bibliography 51
List of Software and Toolkits 56

1 Introduction

In computer science, automatic speech recognition (ASR) signifies the automatic transcription of speech to text and is used in technologies which are present in everyday life. On mobile devices and online platforms, ASR is frequently used to transcribe speech input or implement live subtitling with a level of accuracy comparable to humans. This accuracy is achievable and state-of-the-art for high-resource domains of languages such as English and German. Consequently, many users of ASR have come to expect this level of quality. ASR can also fulfill an important function of making audio data accessible and searchable, for example in broadcast and archives. However, for some languages and domains there is much less training data, which leads to deteriorations of the ASR performance far below human level accuracy. For this reason, ASR is not considered a solved problem. Among the highly problematic cases for current ASR systems is dialectal speech recognition in low-resource environments, resulting in high word error rates (WER): As Ali et al. state, "[for dialectal ASR] the WER can easily go over 40%" (2017, p. 141). This also applies to the ASR model developed at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and included in the Audio Mining system (Schmidt et al., 2016).

In the German-speaking countries, there is a plethora of dialects. Unlike in regions where the traditional dialects have been displaced by newer forms of vernacular[1] speech, in some regions the traditional dialects are highly relevant to the daily lives of many speakers. This is especially true of the southern German-speaking regions. Among the most frequently used and widespread traditional dialects of German is the group of dialects classified as Bavarian. As of writing this and according to our research, however, no public research has shown ASR systems that offer low word error rates on Bavarian. Moreover, even for the more intensively researched field of ASR for the Swiss German dialects, a dialect group that shares many characteristics with Bavarian, the challenges are seen. As an example, Nigmatulina et al. could reach just below 30% for a Swiss German test set in their best two results (2020). This shows the relevance of research in this low-resource environment. For this reason, our work is dedicated to Bavarian speech recognition and will aim to explore if the ASR performance in this case can be improved via different approaches. Among the evaluated approaches are rule-based lexicon additions and data-driven approaches.

[1] Which in German terminology (*Umgangssprache*) is a variety that does not retain the primary features of the traditional dialect such as phonemes particular to the dialect of the region (Wiesinger, 1980).

1.1 Bavarian Dialect

The term *Bavarian* is used in the sense of the dialect group Bavarian, in German linguistic terminology *Bairisch*. This work is explicitly about the dialectal speech of this traditional dialect. Hence, mere Bavarian accent in Standard German or a vernacular are not considered relevant to this work. We apply both the terms *dialect* and *language* to the variety Bavarian, using a neutral meaning of the term *language* and referring to standardized languages as *standard languages*. Thereby, *dialect* and *language* are not in contrast, *dialect* and *standard language* are in opposition, and all are called *varieties* (of German). Since we use the dialectological definition of *High German*, Bavarian is a High German dialect (see Chapter 2.1 Linguistics).

1.2 Problem Definition

German ASR models are very well suited to the task of transcribing Standard (High) German but fail to transcribe the dialect Bavarian with acceptable levels of accuracy. Whilst this disparity is attributable to the low amount of training data[1] for Bavarian — this factor is only adjustable with great effort in recording and annotation — the issue may be broken down into smaller parts that might be individually solvable and open up new potential via bootstrapping. There are several differences between Standard German and Bavarian. One of the most relevant aspects is pronunciation, which affects the acoustic modeling including the lexicon. The pronunciation differs in several sub-aspects, for which historic sound changes are a relevant cause. The phoneme inventory is different, which may create challenging situations when Bavarian sounds occur which Standard German lacks (e.g., front vowel /a/, short close /e/ (Zehetner, 1985, pp. 15,82)). Secondly, related words may differ due to sound changes, even though all relevant sounds are shared between the languages (e.g., *schon* vs. *scho*). Thirdly, the (arguably) shared phonemes differ in their concrete pronunciation (*phones*) or react to their phonetic environment (surrounding sounds) in different ways, which will result in the acoustic model assigning different probabilities to the internal sound representations.

Another aspect in which Bavarian and Standard German differ is their lexicon, which can surface in different meanings and uses of the same words (e.g., *Schwamm* 'sponge'[2] or 'mushroom') or different words being used that are not predictable via sound correspondences (e.g., *Brötchen, Semmel* 'bread roll'). This affects the ASR lexicon and language model.

[1] Ideally, "natural training data", data used in other contexts than linguistics or ASR, is preferable.

[2] This linguistic notation convention may be read as: *the word Schwamm means sponge.* It uses italic letters for object-language and metalinguistic usages of a term (e.g., examples or terminology). Single quotation marks are used for translations. (Max Planck Institute for Evolutionary Anthropology, 2014)

Lastly, there are several differences in the aspects that are modeled by language models. We might distinguish syntactical aspects, concerning sentence formation, and morphological aspects, concerning word formation including inflection.

1.3 Relevance

The current state of ASR is characterized by human-level performance in high resource environments, yet in low-resource environments the performance degrades. For Bavarian as for other dialects — even for the particularly vital dialects of Swiss German — there is little training data. Hence, the performance issue has not been solved for Bavarian and other German dialects. As stated by Ali et al. (2017), we may encounter word error rates above 40% which is supported by our own findings: On our large Bavarian test set, the Fraunhofer IAIS hybrid model achieves a WER of 52.96%. Reaching a WER below 40% has a particularly high significance according to Stadtschnitzer et al. who state that even in spontaneous dialectal speech in noisy conditions, WERs below 40% are an important requirement for a modern ASR system "especially when the ASR output is further processed for tasks like named entity recognition" (2014, p. 3887). In 2015, Stadtschnitzer and Schmidt even found error rates as high as 90.1% on a Bavarian test set (2015).

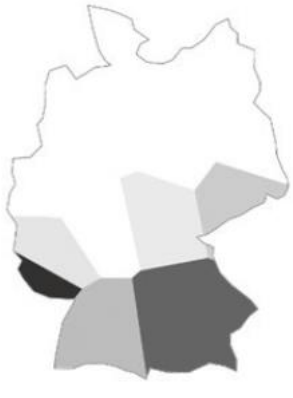
While many traditional dialects have lost most of their significance and number of speakers to Standard German, this is not true for Bavarian. The number of speakers for Bavarian is not as simply determined as for standardized languages. However, in surveys from the last decades, it seems clear that even considering Germany alone, Bavarian is among the liveliest (Adler & Plewnia, 2020; Zehetner, 1985, pp. 155-157).

Figures 1 and 2, taken from a complex figure (Adler & Plewnia, 2020)[1] show that in a survey from 2017, dialect speakers from Bavaria (low right) commonly replied to speak no "[Standard] High German", or to only speak dialect. Only the small federate state "Saarland" was more dialectal by this measure.

Figure 1 Only "High German" [1] *(Adler & Plewnia, 2020)*



Figure 2 Only "Dialect" [1] *(Adler & Plewnia, 2020)*



While Bavarian is not the only dialect group in Bavaria, Bavaria is neither the only state where Bavarian is spoken. Most notably the vast majority of the dialects in Austria are linguistically seen as belonging to this same

[1] The coloring of the polygons corresponds to the mean of the replied numbers; the higher the mean, the darker the polygon is colored (ibid.).

group of Bavarian (Zehetner, 1985, p. 16). If not only Bavarian is wide-spread but also a large portion of the speakers do not actively use Standard German, there is a need for functioning Bavarian ASR. Possible examples for such use cases are subtitling for regional Bavarian broadcast, ASR on smart devices, and automatic tagging or subtitling for large archives (cf. Gref, 2022; cf. Schmidt et al., 2016).

1.4 Objective and Research Questions

Bavarian speech recognition in this work is designed to take Bavarian audio as input and transcribe it to Standard German text, as it was done in similar work for Swiss German, for instance by Nigmatulina et al. (2020) and Stadtschnitzer and Schmidt (2018). For this goal, existing Standard German ASR models are modified which enables us to benefit from the similarity between the varieties.

We consider both Wiesinger's (1980) "Basisdialekt" (the most rural form of dialect and minimally influenced by other forms of speech) and "Verkehrsdialekt" (dialectal varieties with a relevance to a larger area) as conceptually relevant to this work, with ASR for "Verkehrsdialekt" being the realistic goal, as these are the expected data in most of our data sources. They are intelligible to more Bavarian speakers than the most archaic rural forms of Bavarian, whose distinguishing characteristics are limited to regions as small as villages. We argue that the language variety this work is most applicable to is a form of Central Bavarian, oriented towards the dialect of the political center Munich and widely intelligible to Bavarian speakers. In the context of this terminology, the Fraunhofer IAIS hybrid system may be considered mainly oriented towards a standard language and vernacular, "Umgangssprache" in Wiesinger's terminology (ibid.).

This is an explorative work, since the word error rates are notably high, and the training data and the number of previous research projects is limited. Our objective is to explore the options how Bavarian Speech recognition can be improved, taking existing Standard German speech recognition as a basis.

We use the term *rule-based* specifically for handcrafted phonetic mapping rules based on linguistic knowledge, which is opposed to the data-driven approaches in this work that use data to train the statistic models.

Research Questions

Q1: *How can Bavarian ASR be improved on Standard German ASR systems?*

German ASR systems are specialized on Standard German and achieve high word error rates on Bavarian data. The issue posed by Bavarian is resource scarcity, which conflicts with the common training strategy. We therefore aim to evaluate different strategies to improving the performance of existing German ASR systems. These will be rule-based or data-driven.

Q2: *How can the limited amount of Bavarian data be utilized effectively for ASR model creation and evaluation?*
Even if pure rule-based approaches are conducted, transcribed speech data is necessary to evaluate the performance of ASR models. In the case of data-driven approaches, such data is also necessary for training. Due to the scarcity of this kind of data for Bavarian, careful consideration is necessary about how the collected data can be split into training, validation, and test sets effectively.

Q3: *Can a model be created that can reduce the WER below the 40% mark?*
As Stadtschnitzer et al. argue, an ASR model should achieve word error rates below 40% even in the most challenging circumstances (2014). However, we observe higher word error rates on average for our Bavarian data. We will determine if the approaches in this work can achieve an average of under 40% WER for our Bavarian data.

1.5 Limitations

In this last section of the introduction to our work, we would like to state a few limitations to this research:

- This research is conducted by modifying existing Standard German ASR models.
- There is a continuous spectrum between the dialect and the standard language. It is to be expected that the data will contain speech that must be classified as less dialectal or even non-dialectal. This cannot be prevented and must be considered when viewing our results.
- Whilst rule-based approaches regarding n-grams may be conceivable, the rule-based approach in this work was limited to handcrafted phonetic mapping rules.
- As of writing this, the code, the models, and the Bavarian data cannot be shared due to usage rights limitations. We provide the information about our sources and our procedure, which may be used to reproduce our research.

2 Fundamentals

In this section, we will give an overview of the relevant fundamentals in Linguistics and ASR, followed by an overview of our literature review specific to the problem of dialectal speech recognition.

2.1 Linguistics

This section on the linguistic background of this work focuses on concepts from phonetics, phonology, and German dialectology due to their relevance to this work.

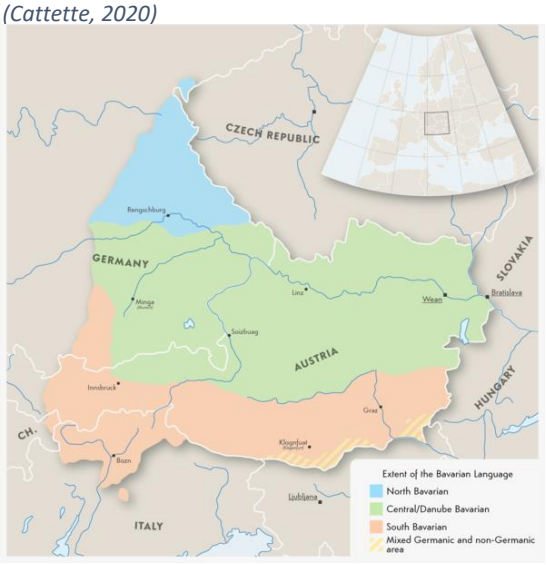
The terms *phone* and *phoneme* are widely used in ASR research and linguistics. While both are used interchangeably in ASR research, they are used for different concepts in linguistics that will be described in short. *Phonemes,* following the definition from structural linguistics, are understood to be the smallest meaning-distinguishing unit in a given language (Uppstad & Tønnessen, 2010; Trubetzkoy, 1939, pp. 32–34) and contrast with the non-distinctive units called *phones*. We will use this widespread concept here but will mention that generative phonology focuses on a concept called *underlying representation* (Chomsky & Halle, 1968, pp. 11–12). A single phoneme of the language replaced by another phoneme may result in a different word: e.g., *Phon* /fon/ vs. *Mohn* /mon/. These two words will be called *minimal pairs* due to the minimal distinction. The set of phonemes which is always language-specific and their phonological modification rules (according to which the abstract *phonemes* assume surface forms called *allophones*) make up the phonology of the language. Phones have no requirement as to distinguishing meaning. Phonological notation is set in slashes / / while phonetic notation is set in (square) brackets [ ]. (Furhop & Peters, 2013, Sections 1.2–1.3)

To distinguish the phonetic variables internal to our hybrid ASR model from our phonetic transcriptions in square brackets [ ], we chose to use double brackets (*white square brackets*) ⟦ ⟧. For more information see Section 4.1 Rule-Based Lexicon Additions.

Dialectology:
Wiesinger distinguishes *Schriftsprache* ('written language') from *Standardsprache* ('standard language') (1980). In this work, we use the term standard language as encompassing both. *Hochdeutsch* 'High German' is a term that is (mostly outside of linguistics) frequently used for a variety of German that is perceived as the standard language (Pittner, 2013, p. 15) and for which we will use the term *Standard (High) German* instead. Nevertheless, the term *High German* has its origins in the geographic differences between the high, mountainous south of the German-speaking region and the low, flat northern regions (ibid.). The dialects from the "high" areas, including Bavarian, are characterized by the sound change *High German consonant shift* (compare the consonant differences in the cognates between English *water, apple, make* and German *Wasser, Apfel, machen*). Consequently, *High German* is a term that applies to the southern dialects that also gave rise to Standard High German. Therefore, unlike in some non-linguistic usages, our usage of *High German* always includes the dialects that underwent the High German consonant shift. Such sound changes are one of the main ways to categorize the dialects and are significant to rule-based approaches of this work. As one of the most southern German dialect groups collectively called *Upper German* (cf. brown area in

Figure 4), Bavarian is characterized by having completed the High German consonant shift. (Niebaum & Macha, 2014, pp. 9, 250–252)

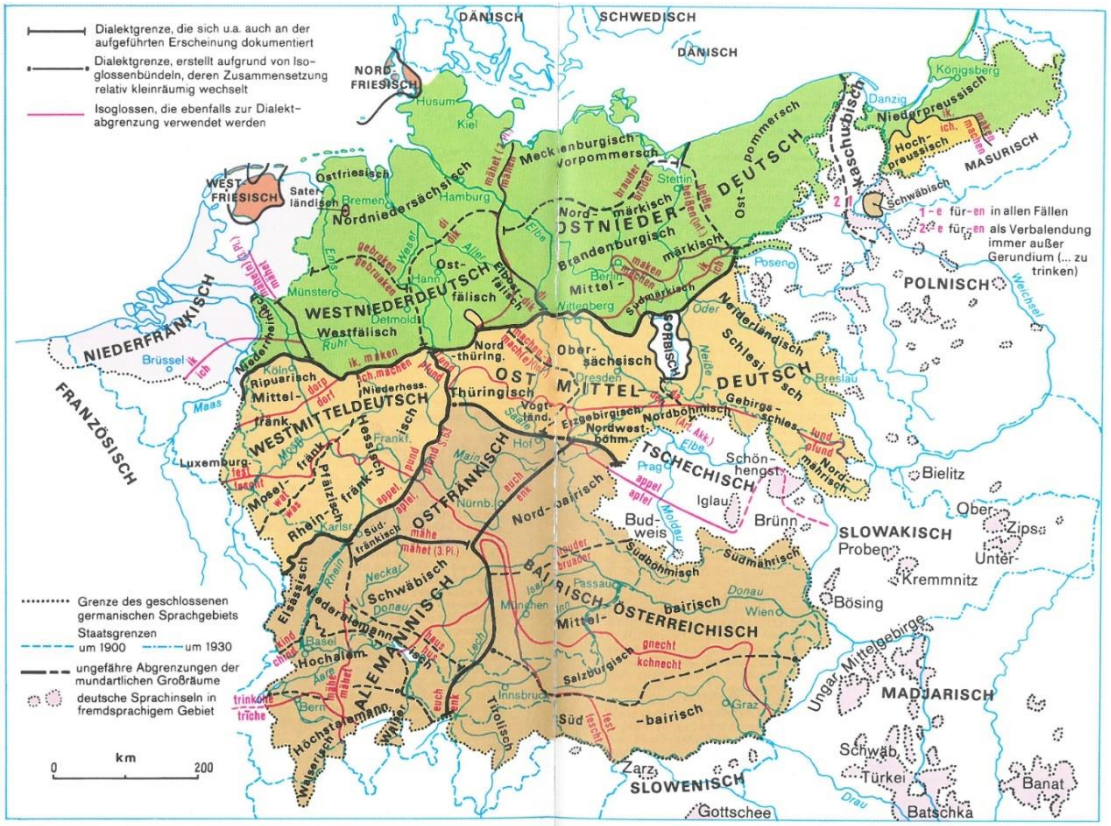
Figure 3 North, Central and South Bavarian area (Cattette, 2020)



Bavarian as a dialect group is native not only to Germany's southern federate state Bavaria but also (from a linguistic standpoint) in most of Austria. Bavarian is categorized into three main dialects: Northern Bavarian, Central Bavarian and Southern Bavarian (Niebaum & Macha, 2014, p. 253; Zehetner, 1985, pp. 60,61). These can be seen in Figure 3.

In Figure 4, all German dialects can be seen as they were distributed in 1900. It shows the isoglosses, the borders of dialect features and thereby dialects, and as can be seen

Figure 4 Central European dialects of Germanic descent, state of 1900 (König, 2011)



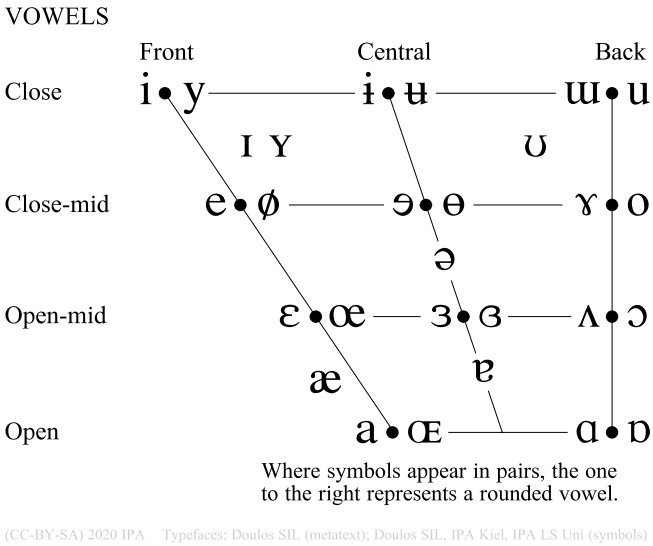
here, usually determined by the distribution of sound changes. The relevant dialect group called *Upper German* is marked in brown. In the orange and green areas north of the Upper German area, the German words *Apfel, Pfund* have a simple plosive sound like English *apple, pound*. North of Central Bavarian — which preserves the Middle High German (mhg.) opening diphthong in *Bruader* — the bruader-brouder isogloss marks the linguistic border to North

Bavarian. To the south at the border to South Bavarian, it shows the isogloss marking the final stage of the High German consonant shift: *kchnecht* corresponding to Standard German *Knecht.* (Niebaum & Macha, 2014, pp. 252–253,162; Zehetner, 1985, pp. 60–69)

2.2 Phonetic Alphabets

In order to have a means of transcribing precisely how words are pronounced, phonetic alphabets are used. Phonetic alphabets are used in linguistics and for ASR lexica. A notable example is the "International Phonetic Alphabet" (IPA) (International Phonetic Association, 2015). For ASR lexica, special characters are often avoided by basing the phonetic alphabet on ASCII characters. An example of this is SAMPA[1] (Wells, 1997). When referring to the hybrid ASR model, we will use the subset of BasSAMPA[2] employed in the Fraunhofer IAIS hybrid model. Figure 5 shows the vowel diagram of the IPA:

Figure 5 IPA vowel chart (IPA Chart, 2020)



The position of the characters approximately resembles the position of the tongue. Many pairs of rounded and unrounded vowels are shown: The right vowel in each pair is pronounced with rounded lips while the other is pronounced with unrounded lips. (International Phonetic Association, 2015, p. 10).

Table 1 lists relevant phones and phonemes with their transcription and a description (without absolute specificity) — the focus of the description is to identify features and not uniquely identify or define the phone. Furhop and Peters (2013, Sections 1.2–1.3) provide an overview of the German phonology. However, in our transcriptions in the column "German

[1] www.phon.ucl.ac.uk/home/sampa

[2] https://www.phonetik.uni-muenchen.de/Bas/BasSAMPA

Examples" there may be some differences to their transcriptions. We chose the words and transcribed them as we deemed it most accurate but do not wish to imply that our transcription is optimal. While we aim to provide German phonemes when appropriate (e.g., /ɛ/), we only provide phones in ambiguous cases to avoid implying phonemicity (e.g., [ŋ]). To illustrate multiple pronunciations of one phoneme we supply several phones to one phoneme (e.g., /e/, [e] - [eː]).

Table 1 BasSAMPA and IPA

BasSAMPA	**IPA**	**German Examples**	**Description**
[e]	[e]	/e/ [e] *Beton* (unstressed)	A short close-mid front vowel
[e:]	[eː]	/e/ [eː] *See* (stressed)	Long [e]
[E] [9] [O]	[ɛ] [œ] [ɔ]	/ɛ/ *[*ɛ] *Welle* /œ/ [œ] *Völle* /ɔ/ [ɔ] *Wolle*	Open-mid vowels
[I] [Y] [U]	[ɪ] [ʏ] [ʊ]	/ɪ/ [ɪ] *stick'* /ʏ/ [ʏ] *Stück* /ʊ/ [ʊ] *Stuck*	Lax (near-) close vowels
[2] [9]	[ø] [œ]	/ø/ [øː] *Höhle* /œ/ [œ] *Hölle*	Two rounded front vowels corresponding to German *ö*
[@] [6]	[ə] [ɐ]	[ə] *äh* (unstressed) [ɐ] *der* (unstressed)	Reduction vowels in German
[aI] [aU] [OY]	[a͡ɪ] [a͡ʊ] [ɔ͡ʏ]	[a͡ɪ] *Mais* [a͡ʊ] *Maus* [ɔ͡ʏ] *Mäuse*	Diphthongs in BasSAMPA
[C] [x]	[ç] [x]	/ç/ [ç] *ich* /ç/ [x ~ χ ~ r̊χ] *Tuch* /ç/ [χ ~ r̊χ] *mache*	Dorsal (articulated with the "back" of the tongue) fricatives
[N]	[ŋ]	[ŋ] *Wange*	velar nasal
[Q]	[ʔ]	(*be-*)[ʔ]*achten*	Glottal plosive/ stop

While BasSAMPA has a character for stress, our lexicon does not employ it, which makes predictable allophone[1] alternations in German unpredictable. Now, phonologically predictable features of German like the glottal plosive [ʔ] and vocalic length in pairs such as [e], [eː] (Furhop & Peters, 2013, Section 1.3) can resemble minimal pairs in the ASR lexicon and might have a meaning-distinguishing function. The lack of a marker for syllable borders is also notable.

2.3 Features of Bavarian

In this section, we will give an overview of the characteristics of some Bavarian and Standard German sound changes and features since they are highly relevant to the acoustic modeling of our ASR models. We argue that in this work, Central Bavarian is most significant, while Southern Bavarian (cf. p.7 Figure 3) is mostly spoken outside of Germany (Zehetner, 1985, pp. 62,63) and therefore relevant for possible future research that takes Austrian data sources into account. As far as we can determine, our audio data almost exclusively reflects Central Bavarian as opposed to Northern Bavarian (cf. p. 7: Figure 4), so Central Bavarian will be the focus of this section.

The following information on Bavarian is described by Zehetner (1985, pp. 16,25,44,54,56,84) and Niebaum and Macha (2014, pp. 252–253). Bavarian is classified as Upper German since the High German consonant shift also affected words such as *Apfel* and *Pfund*, where northern dialects did not shift the original plosive sound (see 2.1 Linguistics). Disregarding South Bavarian, this is the same stage of the sound change which applies to Standard High German. Another shared feature between these two varieties is that the original long vowels mhg. *î*, *û* and *iu* became diphthongs such as in *weiß*, and *Maus*. Apart from such closing diphthongs (ending in a closer sound), Central and South Bavarian also retain the opening diphthong series that was shifted to long close vowels in Standard German:

Bruada - Bruder, *Liab - Liebe*, *miad - müde*

While the new diphthong series (*weiß, Maus*) exists in both varieties, only Standard German merged the new diphthongs (as in *Wein*, *Säure*, *sauer*) completely with an old diphthong series (as in *Bein*, *Stäube*, *Staub*): Standard German has identical pronunciations of the words *weiß* ‘white’ from the word *weiß* ‘knows’. For the old diphthong in *weiß* ‘knows’, a widespread sound change led to the common form *woaß* in many Bavarian sub-dialects and thereby another opening diphthong.

[1] Phones that realize the same phoneme are called *allophones* of the phoneme (Furhop & Peters, 2013, p. 43).

While this overview cannot describe Bavarian completely, the following sound changes will also be highly relevant to this work:

- The sound change *umlaut* led to rounded front vowels in German, such as <ö>[1] /ø/. These were unrounded in Bavarian, merging with the corresponding unrounded sound (e.g., /ø/ - /e/ > /e/, *bösen* > *besen*).
- The distinction between the sounds /p/ - /b/, /t/ - /d/ and /k/ - /g/ is largely lost in Bavarian; unaspirated, unvoiced plosives are common.
- Some sound changes affected final sounds. Most notably, in Middle Bavarian, /l/ was lost in final position, becoming a front vowel and causing further sound changes (e.g., *vil* > *vui*).

2.4 ASR Fundamentals

In ASR, the term *phoneme* is used without distinguishing it from the term *phone* (Moore & Skidmore, 2019). Due to the significance of linguistic concepts for the linguistic rule-based approach in this work, we will distinguish the terms. Outside of the ASR system, the terms are used as defined in (2.1 Linguistics). A different usage of these terms when referring to the ASR system would require inconsistencies (see Section 4.1 Rule-Based Lexicon Additions). We therefore had two options: Using a third neutral term[2] or applying the linguistic terminology to the ASR system. Even though the situation for the ASR system may be different than for a real speaker, we determine that the transcription in the lexicon clearly is designed based on German (allo-)phones and not phonemes. We therefore will use the term *phone* in this chapter where other research uses *phoneme*.

As can be seen from the paradigms statistical (cf. Stadtschnitzer, 2018) and end-to-end (cf. wav2vec 2.0, cf. Baevski et al., 2020) ASR, the common approach is data-driven ASR. A complementary approach is knowledge-based, or rule-based ASR (Sushmita, Vijayshri, & Challa, 2016). This work is based on data-driven ASR, since statistical and end-to-end models are used. In this work, rule-based approaches are applied to a statistical model, which must be distinguished from a knowledge-based implementation of an ASR system. The rule-based aspect of this work must be understood as handcrafted linguistic rules that are applied to the existing lexicon data.

---

[1] Like [ ] for phones, / / for phonemes, < > are used to signify graphemes and grapheme sequences (Max Planck Institute for Evolutionary Anthropology, 2014). The tilde (~) is placed between alternatives.

[2] If this work had a chance of coining such a term, we would suggest *phonoid (-oid* 'resembling; having the likeness of' https://en.wiktionary.org/wiki/-oid). While we couldn't find a similar usage of this term, the idea may have ultimately originated from Pike's terms *vocoid* and *contoid* (1943).

Statistical ASR

The following description of statistical ASR is based on Stadtschnitzer (2018, pp. 21–30) if not cited differently. One of the most successful ASR approaches are statistical ASR systems which are defined as being based on the Bayes' theorem:

$$P(w_1^N|x_1^T) = \frac{P(x_1^T|w_1^N) \cdot P(w_1^N)}{P(x_1^T)}$$

The Bayes' theorem, as it is done in the above formula, can be applied to ASR for the probability of a word sequence[1] w_1^N given a sequence of acoustic features x_1^T resulting from feature extraction. In other words, it calculates the probability of a predicted transcription given the speech input. Since the intended output of an ASR system is the most likely word sequence given the input, the shared factor $p(x_1^T)^{-1}$ can be disregarded (Stadtschnitzer, 2018, p. 22) and the problem can be mathematically defined as:

$$\begin{aligned}[w_1^N]_{opt} &= \underset{w_1^N}{argmax}\{P(w_1^N|x_1^T)\} \\ &= \underset{w_1^N}{argmax}\{P(x_1^T|w_1^N) \cdot P(w_1^N)\}\end{aligned}$$

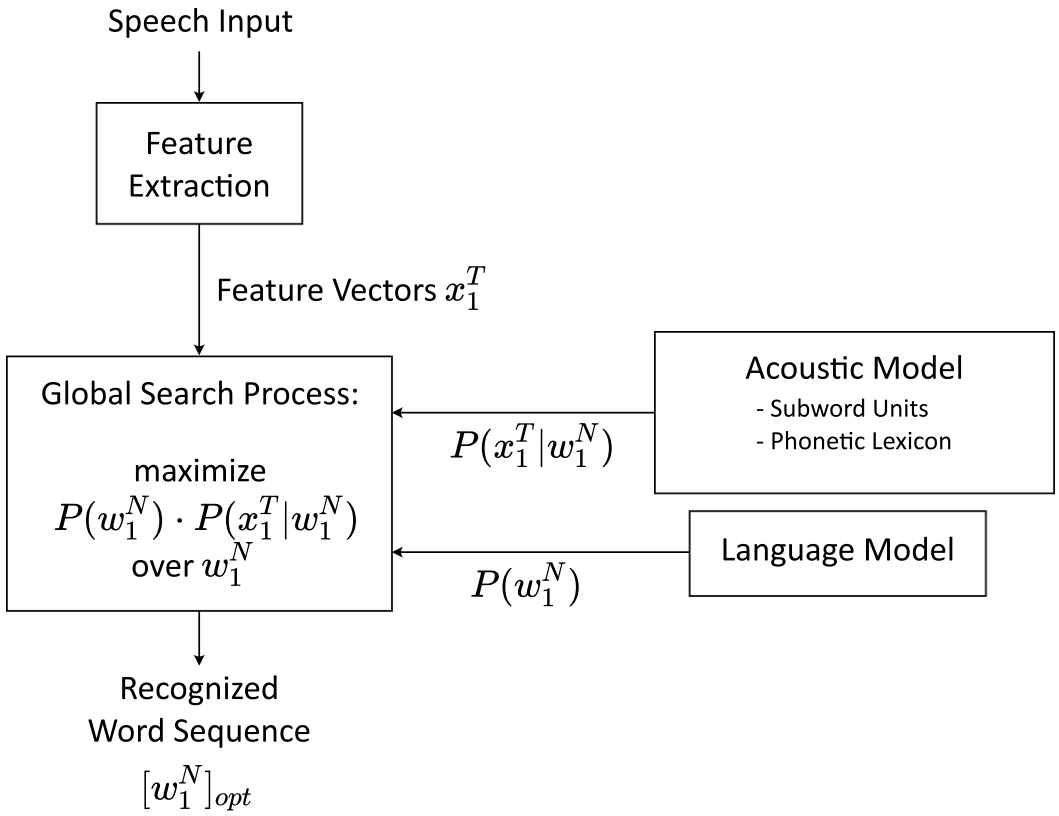
Thus, the a priori probability of a word sequence $P(w_1^N)$ can be calculated independently of the acoustic probabilities $P(x_1^T|w_1^N)$. This way, an ASR system can be implemented using an acoustic[2] model (AM) component and a language model (LM) component that can be trained separately. These statistical models are implemented successfully using an AM architecture known as *HMM-based.* Hidden Markov models (HMMs) realized via Gaussian Mixture Models (GMMs) were used with good performance, but even better results could be achieved when the GMM component was replaced with deep neural networks (DNNs) due to their better generalization (Gref, 2022, p. 4). These are called *hybrid ASR models* (ibid.) and are the kind of statistical ASR model used in this work. We use the terms *hybrid ASR model* and *hybrid model* interchangeably since the context of this work is ASR.

[1] We use the same notation for sequences as Stadtschnitzer (2018, p. 21) and Ney & Ortmanns (1999) to increase the readability: w_1^N as opposed to the more traditional notation $(w_n)_{n=1}^N$.

[2] Stadtschnitzer uses the term "acoustical model" (2018, p. 25), however, we adopt the term used by Gref (2022, p. 15).

Figure 6 gives an overview of a statistical ASR system. To the right, the independence of the AM (including the lexicon) and the LM is shown. After this short description of the components of statistical ASR, we will describe the setup of such systems in more detail, beginning with the speech input.

Figure 6 Overview of a statistical ASR system adapted from Ney (1990)



The infinite number of possible pronunciations of speech share acoustic patterns that enable the distinction of a finite number of sounds. Instead of inputting a raw wave form of speech, the input can be pre-processed into a form that is more beneficial to the intended task via a process called *Feature Extraction*. The discriminative features depend on the task: In ASR, representations of the spectrogram are beneficial compared to wave forms. Examples of widely used discriminative features in ASR are mel-frequency cepstral coefficients (MFCCs) and filterbank coefficients. (Stadtschnitzer, 2018, pp. 11,12)

The extracted discriminative features are passed to the AM that stochastically models "the temporal and static features of the speech signal" (Stadtschnitzer, 2018, p. 25). Its tangible function is modeling the probability of phones given the acoustic data. This is commonly achieved via hidden Markov models (HMMs), statistical models realized by a graph of unob-

servable (hidden) states. Since the pronunciation of phone(me)s[1] varies due to the phones in their environment (cf. 2.1 Linguistics), context dependent phones can be used to model this. Triphones have the advantage that they model the pronunciation of a phone in relation to sounds before and after it. On the other hand, the number of possible triphones is so high in many languages that some triphones occur rarely in training if at all; but this can be alleviated by clustering similar observation probability distributions. (Stadtschnitzer, 2018, p. 25)

Statistical ASR models have a list of all words that can be recognized. Apart from a word's spelling, it lists its pronunciation via a phonetic transcription (cf. 2.2 Phonetic Alphabets). This word list is called a *pronunciation dictionary* or *lexicon*, as we call it in this work. In the case of homophony, several words in the lexicon can be supplied with the same transcription; also, several pronunciations for the same word are possible. Since such a lexicon can have millions of word forms, manually transcribing all of them is not feasible. Therefore, grapheme-to-phoneme (G2P) conversion models are used that can be trained with known pronunciations. After training, they can generate new pronunciations to words that were never encountered during training. (Stadtschnitzer, 2018, pp. 23,24)

Independently of the acoustic probabilities, an LM in ASR predicts the probability of word sequences and thereby models the grammar and semantics of the language statistically. A widely used type of LM is the n-gram model (Stadtschnitzer, 2018, p. 27). The n-gram model (Bahl et al., 1983) calculates the probability of a word sequence of n words using the chain rule, which allows the probability to be calculated iteratively using the probability of each word given its predecessors. This can be seen in the following formula.

$$P(w_1^M) = P(w_1, \dots, w_M) = \prod_{m=1}^{M} P(w_m | w_{m-1}, \dots, w_1)$$

$$\approx \prod_{m=1}^{M} P\left(w_m | w_{m-1}, \dots, w_{m-(n-1)}\right)$$

Note that the n-grams must be shortened at the beginning.

The above formula shows the probability of a word sequence of M words in the first line. As the second line shows, an n-gram language model approximates this probability by only considering up to $n-1$ predecessors of each of the M words. While in human langua-

[1] This is influenced by the transcription system used in a given ASR model. While we argue that the Fraunhofer IAIS hybrid model uses a somewhat broad phonetic transcription (phones), other models may use completely phonemic transcriptions (phonemes) or particularly narrow phonetic transcriptions that differentiate between very subtle differences. However, even a narrow transcription cannot capture all possible differences.

ge grammatical rules are theoretically unbounded by the number of words they span[1], such a model with a restricted n performs well for ASR purposes. (Stadtschnitzer, 2018, pp. 27,28)

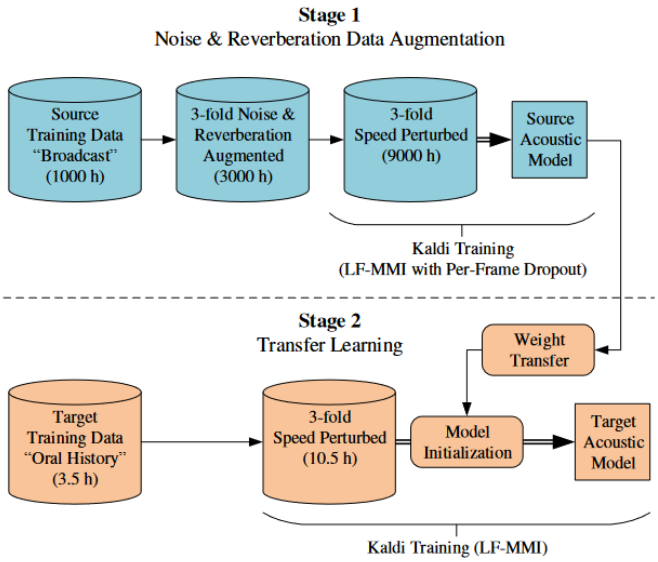
Transfer Learning

As described in this work, low amounts of data are a great challenge when trying to implement data-driven speech recognition. This is especially significant if a model must be trained from scratch, which takes a great amount of time and resources (Gref, 2022, p. 133). Therefore, it is sought to re-use information learned by ASR systems in new ASR systems for new tasks and settings via a training paradigm called transfer learning (TL). "The initial idea of transfer learning is to reuse the experience/knowledge obtained already to enhance learning for new things." (Wang & Zheng, 2015). There are various approaches that may be categorized as such. Wang and Zheng name "multitask learning, lifelong learning, knowledge transfer, knowledge consolidation, model adaptation, concept drift, covariance shift" (2015). However, the categorization varies with the author (ibid.). In the case of ASR "[t]he basic idea is that the features learned by DNN models tend to be language-independent at low layers and more language-dependent at high layers" (ibid.).

Gref et al. used this approach to improve ASR for a low-resource domain (2019). TL is applied in Stage 2 of Figure 7 where they re-used the DNN-weights obtained from training in a high-resource domain for a recognizer specialized for the low-resource domain of oral history interviews.

This is not only applicable to different domains and dialects of one language but also to multiple languages, as shown by Baevski et al. (2020) in the end-to-end system wav2vec 2.0.

Figure 7 Data Augmentation and TL (Gref et al., 2019)



[1] (cf. dangling prepositions or German *Verbalklammer*)

Wav2vec 2.0

In the field of low-resource ASR, an important development was finding a way to pre-train DNNs in a language-agnostic way so that language specific recognition can be effectively trained with the low data available for many languages. For this purpose, Baevski et al. propose wav2vec 2.0, a transformer-based model with an unsupervised pre-training task (2020). This end-to-end approach — using the knowledge gained from an unsupervised high-resource task to better learn a low-resource task — can be seen as an example of transfer learning, adhering to Wang and Zheng's description of the term (2015). End-to-end models, unlike hybrid ASR models, do not belong to the statistical ASR approach that uses a pipeline of models as described in this chapter; instead, they use a single model to map audio to letters in a single task (Baevski et al., 2020; Gref, 2022, pp. 28,29).

2.5 Low-Resource and Dialectal ASR

We observe that there is little published research on Bavarian ASR. However, research regarding Swiss German can be considered particularly relevant to this work due to the similarities between the dialects.

One of the findings in the research on Swiss German ASR that are most relevant to this work was made by Nigmatulina et al. They found that in ASR systems, transcribing the spoken dialect to orthographic Standard German is not only viable, but even advantageous (2020). This information is crucial since it directly influences the choice of data for this problem and allows rule-based modifications to the existing Standard German model. Our work aims to achieve such transcription into Standard German from Bavarian.

Garner et al. worked on dialectal speech recognition for Walliser German and found supporting evidence for their hypothesis that bootstrapping a Standard German model to this High German dialect is effective (2014). Their argumentation supports the approach of modifying a Standard German model.

The most relevant research on rule-based lexicon additions that we could find was conducted by Wester and Kessens (2000) for Dutch pronunciation variation in ASR, not least because of the close relationship between Dutch and German. Their results support the rule-based approach.

2.6 Methodology

In this section we will go into detail about the methodological aspects of our work, such as the technological background and the metrics used for our experiments.

2.6.1 Standard German Baseline Models

In this section, we will present the models this work tries to improve. While this is a short abstract description, the fundamental concepts are described in (2.4 ASR Fundamentals).

Hybrid model

We begin by describing the hybrid model used in the Fraunhofer IAIS Audio Mining System (Schmidt et al., 2016), as it is described by Schmidt (ibid.) and Gref (2022, pp. 71,126–127,201–207). These works may be consulted for further information. The Fraunhofer IAIS ASR model is a hybrid DNN-HMM ASR model, an HMM-based model created using the Kaldi toolkit (Povey et al., 2011). It is optimized for and trained on Standard High German TV and radio data. Its AM, the 3-fold v2, is more precisely a LSTM-TDNN (Peddinti et al., 2018), a time-delay neural network combined with long short-term memory layers and trained on LF-MMI (Povey et al., 2016), which means it is trained in a sequence-discriminative training procedure. Three LSTM layers with a cell dimension of 1024 and a projection dimension of 256 and seven 1024-dimensional TDNN layers are used. It uses MFCCs for phone recognition and i-vectors for speaker recognition as feature input: “a 300-dimensional input at each time-step consisting of five consecutive 40-dimensional MFCC features and a 100-dimensional i-vector” (Gref et al., 2019).

The hybrid system uses a 5-gram LM trained on German broadcast text data and has millions of words. For the creation of the lexicon, Sequitur G2P[1] (Bisani & Ney, 2008) and the pronunciation database PHONOLEX[2] were used. All components are combined in a weighted finite state transducer (WFST) on which graph search is conducted to find the most likely word sequence for the input.

Since the Fraunhofer IAIS model is trained every day with current data, we used an older fixed model version as a baseline model (german-default-1.0.124.20201101, created 2020-11-01) whose AM, however, is still current. The lexicon and LM of this version have approximately 3,500,000 words**.** This hybrid baseline model was also used by Pritzen (2021, p. 36). It will be called *hybrid baseline* in this work.

Wav2vec 2.0

Apart from the state-of-the-art hybrid model for Standard German, an End-to-End approach

[1] https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[2] https://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html

was conducted. For this, the framework wav2vec 2.0 (Baevski et al., 2020) was chosen due to the advances its pre-training procedure offers for all languages and especially low-resource languages. We used two models that are fine-tuned to Standard German and based on Facebook's wav2vec2-large-xlsr-53 (Conneau et al., 2021).

The architecture of the model *wav2vec2-large-xlsr-53* corresponds to *wav2vec 2.0 large* which has 317 million parameters (Baevski et al., 2020). It was pre-trained on 53 languages, amounting to 56,000 hours of data, and via the XLSR procedure (Conneau et al., 2021).

We considered two wav2vec 2.0 models for testing and fine-tuning on Bavarian. Previous work at Fraunhofer IAIS resulted in a German wav2vec 2.0 model fine-tuned on broadcast data. This model was used in published work by Pritzen (2021, p. 36) and will be called *internal-IAIS-wav2vec2* in this work. One of the advantages of this model is that it was fine-tuned on the GER-TV1000h corpus (Stadtschnitzer et al., 2014), which suggests that it is more similar to the Fraunhofer IAIS hybrid model than other models. It is based on a model fine-tuned by Grosman[1], itself a XLSR-53 large model fine-tuned for German using the German Common Voice dataset (Ardila et al., 2020). We will refer to it as *grosman2021-wav2vec2*. Comparing both models provides the advantage that grosman2021-wav2vec2 is a public model, which offers more opportunities for reproduction of our findings. It will also provide some evidence on the influence of more German fine-tuning using GER-TV1000h on the performance on Bavarian speech.

2.6.2 Benchmarking and Metrics

The definitions and formulas in this section are taken from (Stadtschnitzer, 2018, pp. 30-33). For the hybrid models, the benchmarking will be conducted using the Fraunhofer IAIS benchmarking workflow. For wav2vec 2.0, we used jiWER[2] to compare the results to the ground truth. Both benchmarking procedures provide the results in the measure *word error rate* (*WER*), using a case-insensitive definition. The WER calculations for both models are, therefore, equivalent. The WER is determined using the Levenshtein distance (Levenshtein, 1966) of the ground truth and predicted sequence on word level. The Levenshtein distance (d_L) is the smallest number of substitutions (S), deletions (D) and insertions (I) required to "change the hypothesis sentence into the reference sentence" (ibid.). With N being the number of words in the segment, the WER can thus be calculated as:

$$WER = \frac{S + D + I}{N} = \frac{d_L}{N}$$

[1] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german

[2] https://github.com/jitsi/jiwer

In this work, if not otherwise stated, we will measure changes in WER relative to the baseline WER. Due to the definition of WER, mathematically negative changes are seen as improvements in this work. The changes will be given as percentages of the baseline WER, which we will calculate as:

$$\frac{WER_{train} - WER_{base}}{WER_{base}} \cdot 100\%$$

For LMs, two other performance measures are used. The best model for the target domains, predicts a random unseen sample (e.g., a test set) from the domain with the highest probability (ibid.). To determine how well the LM predicts a test set, the perplexity is calculated. Low values are preferable because they show that the model predicts the test set well. "Perplexity is the probability estimate assigned to a word sequence w_1^N by a language model, normalised by the number of words N" (Stadtschnitzer, 2018, p. 32):

$$PP := P(w_1^N)^{-\frac{1}{N}} = \left[\prod_1^N p(w_n|w_{n-1}, \dots, w_1)\right]^{-\frac{1}{N}}$$

Naturally, an n-gram language model cannot predict probabilities of word sequences of unknown words. Therefore, we will also use the running OOV (out of vocabulary) rate, which is the ratio of unknown encountered words by the total encountered words:

$$OOV_{run} = \frac{n_{run,unknown}}{n_{run,total}}$$

2.6.3 List of Approaches

To answer the research questions (see 1.4 Objective and Research Questions), we devised and conducted different approaches which will be listed below: One rule-based approach to reduce the needed amount of Bavarian ASR data and data-driven approaches to improve hybrid and end-to-end ASR models. More information on the approaches will be given in Chapter 4 Experiments.

A1: Rule-based phonetic lexicon additions

A2.1.a: DNN-HMM AM Transfer Learning

A2.1.b: Creation of 5-Gram LM from transcriptions

A2.2: Fine-tuning German wav2vec 2.0 models

3 Bavarian Data

In this low-resource language, finding appropriate data for training and testing is challenging and important. Even for rule-based approaches, finding appropriate data for testing is essential. Only this way can ASR improvements be reached and determined. This chapter deals with the data collection and creation of usable ASR datasets.

3.1 Bavarian Data Collection

There is a lack of transcribed speech data for the Bavarian dialect, so the collection of data was a crucial step. This section describes the Bavarian data we have gathered. We divide the Bavarian language data that was used in this work into four groups:

Oachkatzl-2020-05: Transcribed Bavarian broadcast data supplied by the Bavarian broadcaster Bayerischer Rundfunk (BR). The largest source of these audio files is the television series “Dahoam is Dahoam”. This data has a total length of 33.19 hours.

Bairisch-2020: Transcribed data of varieties of German spoken in Bavaria, including Bavarian which is labeled as *Altb* (compare the Bavarian region “Altbayern”) to contrast it from other dialects. This data was supplied by BR. The relevant segments amount to 0.15 hours.

DSA-Wenker: Linguistic interviews with five speakers of Central Bavarian including transcriptions provided by Deutscher Sprachatlas (DSA), as part of their "Regionalsprache.de" research project[1]. The interview questions follow the catalog of sentences by the linguist Georg Wenker that is designed to concisely gather linguistically relevant information across German dialects (Niebaum & Macha, 2014, pp. 65–67). As of writing this, the usage rights were limited to this work. We determined that the total length of the relevant segments is 0.22 hours.

“Dahoam is Dahoam”: Apart from the segments included in the Oachkatzl-2020-05 dataset, some of the episodes of this series can be accessed online via the BR[2]. They include closed captions for the episodes. Due to mismatched timestamps and imprecise wording, only the subtitles of this dataset could be used. More information will be provided in the following chapter.

[1] “DSA, Akademie-Projekt Regionalsprache.de”.

[2] https://www.br.de/mediathek/

3.2 Creation of Datasets

This section describes our creation of the ASR datasets in this work. It will also describe our considerations and which data could not be used successfully, which may be useful to other researchers. Following the results of the experiments, we will discuss if they could fulfill their intended use (see 5 Conclusion).

Since Standard German text is the intended output, a canonical spelling exists, which is useful for training and improves the intelligibility of the text output. We received two datasets from the regional broadcasting service BR. This dataset will henceforth be referred to as *Oachkatzl-2020-05* or *Oachkatzl* (bav. 'squirrel'). The Oachkatzl dataset with a length of 33.19 hours is our largest dataset for Bavarian. It is comprised of Bavarian audio data from the regional broadcast with Standard German transcriptions. The largest source of the segments in this dataset is the television series *Dahoam is Dahoam* (bav. '(at) home is (at) home') from which we considered to extract additional training data — the reasons why this was unsuccessful will be provided later.

The other dataset we were provided with by the BR is a small dataset for multiple dialect groups in the state Bavaria including Bavarian, here classified as *Altb* which ultimately refers to a region in Bavaria (German *Altbayern*) and the dialects spoken in it (cf. Zehetner, 1985, p. 19) to distinguish them from the other dialects spoken in the state Bavaria (German *Bayern*). We extracted these segments classified as *Altb* to create a test set for this work. This led to a test set "Bairisch-2020-Altb" of only 100 segments, 1672 words and a length of 0.15h. In other words, the test set is so small that we cannot rely on it as our only test set. It was, therefore, necessary to create another test set. It is also notable that the small dataset may be considered quite specialized due to the statistical weight of for example vocabulary related to injury, sports niches or the few speakers' personal experiences.

Two different approaches to creating more testing data were conceived: Searching for more data in this low-resource domain or splitting the larger training set. The first approach led us to extracting episodes of the series "Dahoam is Dahoam" and contacting the linguistic research center "Forschungszentrum DSA". The additional test sets that were created were kept as separate test sets because of their different sizes and domain. The expectation was that joining the test sets would not serve a benefit and might obscure additional information.

Since the dataset Oachkatzl was much larger than the original test set, we split it into three subsets: Oachkatzl-train [23.95 hours], Oachkatzl-dev for validation [3.9 hours] and Oachkatzl-test [5.3 hours]. We observed that in the set Oachkatzl, there are frequently two adjacent segments with a significant overlap over the identical speech and transcription, which would be a significant issue if this occurred regularly between the training and test sets. We sought to avoid this issue by taking the segment IDs into account, which were generated during the creation of batches of segments by the BR (we were informed that these do not

provide information on the source of the data but we found that segments with identical speech and text extracts could indeed be found by checking the following IDs). This is the best solution we could find to create a sufficiently large test set with the information at our disposal. It seems that the chance of overlaps still being an issue is small. The main advantage of the sets extracted from Oachkatzl is their length. There are, however, disadvantages to the data, namely that the main source of the segments is a fictional television series. This can shed doubt on the authenticity of the Bavarian, which is scripted beforehand and needs to be intelligible to a wide range of listeners (cf. *Verkehrsdialekt,* 1.4 Objective and Research Questions), and on the variation in the data for the same reason. Here, having multiple test sets is an advantage because joining the test sets would hardly offset this due to their different lengths.

Responding to our inquiry, the Forschungszentrum DSA kindly provided us with Central Bavarian data and a personal usage right to the author for this project. The provided data is from recordings of the so-called "Wenker Sätze" (ger. 'Wenker Sentences') based on the large-scale German dialect survey initiated by Georg Wenker in the German Empire (cf. Niebaum & Macha 2014, pp. 65–67). These sentences are designed to provide the information most relevant for dialectologists (such as phonological data) in a condensed and consistent way for dialects in Germany as well as Austria and Switzerland (ibid.).

We expected the following advantages of this data: They provide the quality and accuracy of transcription that is to be expected from a professional dialectological survey. The data can be safely considered Bavarian and produced by "competent" native speakers. It was also possible to determine the corresponding town or village of the dialect and limit the recordings to the area we consider most relevant to comparability (the Central Bavarian area). These advantages make the data an essential addition to the ASR testing data and may provide benchmark data that is also meaningful to the scrutinous eyes of dialectologists. Ultimately, this resulted in the test set we call DSA-Wenker with a length of 0.22 hours. However, we recognize several issues that prevent large-scale usage of such linguistic data for ASR purposes. Firstly, the total length of the language data is short. Secondly, all speakers speak the same sentences, which may have some advantages in some cases but also limits the usability of the data as reliable benchmarking data. Also, the translated sentences often depend on dated language. Since the data is transcribed and segmented on word level, we considered to also use simple one-word segments. This could not be achieved due to the precise segmentation necessary in interconnected speech, where the previous segmentation was not always sufficiently precise.

The other promising source of possible training data is the series "Dahoam is Dahoam". Since the series has closed captions, the episodes are already segmented and transcribed for our research. In the end, the segment borders and the transcriptions turned out to be imperfect. Using our baseline acoustic model for forced alignment, we could not improve the

situation. This may be solvable, which would give access to a potentially much larger set of training data than Oachkatzl-2020-05, but it was not possible in the course of this work. A training set that resulted from the inclusion of "Dahoam is Dahoam" data was created: "Bavarian-ASR-train". This training set is currently not considered useful for AM training or end-to-end training but useful for LM training since the issues regarding text-to-audio alignment do not affect LM training.

Conclusion

Finally, we created three test sets and one validation set. This resulted in a smaller training set, which was unavoidable due to the small size of the other two test sets. We also gained information about caption-to-speech alignment limiting the exploitation of a large source of data in "Dahoam is Dahoam". Ultimately, the further experiments in this work will show if this creation of datasets was a success (see 5 Conclusion).

3.3 Overview of Bavarian Datasets

Table 2 shows an overview of the data that was used to create and test the models in this work:

Table 2 Bavarian ASR Datasets

Name	Purpose	Len (h)	AvgLn	Seg	Words	W/Seg	UniqW
Bavarian-ASR-train	LM training	(71.28)	(3.51)	73,105	629,323	8.61	31,098
Oachkatzl-train	Training	23.95	5.98	14,425	221,444	15.35	17,966
Oachkatzl-dev	Validation	3.92	5.71	2,472	36,360	14.71	5,060
Oachkatzl-test	Testing	5.32	6.12	3,133	50,403	16.09	5,864
Bairisch-2020-Altb	Testing	0.15	5.28	100	1,672	16.72	659
DSA-Wenker	Testing	0.22	1.85	427	2,482	5.81	409

Abbreviations: Len: Length, AvgLn: Average Length, Seg: Segments, W/Seg: Words per Segment, UniqW: unique words

4 Experiments

In this chapter, we will evaluate the different rule-based and data-driven approaches chosen for this work. We will first explore the option of rule-based lexicon additions. Taking linguistic knowledge into account, we aim to create a workflow for improving the phonetic modeling for Bavarian. Afterwards, we will experiment with data-driven approaches for refitting German hybrid and wav2vec 2.0 models for Bavarian. For this, we will conduct transfer learning and n-gram training.

4.1 Rule-Based Lexicon Additions

Since the phonetic realization of Bavarian words differs in many ways from their Standard German cognates, the phonetic transcriptions in the hybrid baseline model's lexicon are an imperfect fit for the recognition of Bavarian. The phonetic differences between Bavarian and Standard German are, however, described in detail in linguistic resources. Hence, it seems realistic that Bavarian ASR can be improved without the use of any training data. Due to the low amount of Bavarian training data, we decided to explore this option. Since the term *phoneme* cannot be applied to the lexicon without causing contradictions while adhering to a linguistic definition of the term, we will consider the sound units in the lexicon broad (allo-)phonetic variables (retaining a relatively high abstraction level) and use a slightly modified bracket notation (⟦b aI S p i: l⟧) as well as the term *phone*. A concrete proof for our assumption would depend on taking many different phonological analyses of German into consideration, so, we cannot supply this for time constraints. However, the reader may consider that phones such as the following are distinguished in the lexicon:

⟦o:⟧ vs. ⟦o⟧ vs. ⟦O⟧,

⟦C⟧ vs. ⟦x⟧,

⟦Q o:⟧ vs. ⟦o:⟧,

We use a different character for the phonetic brackets to signify that the transcription represents the phonetic variables of the AM and uses BasSAMPA, while our manual phonetic transcriptions use the IPA.

Experimental Setup and Anticipations

In this experiment, the phonetic lexicon is supplemented with new pronunciations for existing words in a rule-based way. Linguistic knowledge about the phoneme inventory and the phonological history of Bavarian is considered to create handcrafted rules that map phones in the baseline lexicon to other phones "known" to the acoustic model. The baseline lexicon has approximately 3.5 million words with transcriptions. These are the source words and source phones to our rules, mapped to target pronunciations. In this work, we call these rules *phonetic mapping rules* and must stress that they are not phonological rules of Bavarian but handcrafted mapping rules of Bavarian pronunciations to Standard German sounds. The problem how to formulate rules turned out to be more complex than initially assumed. Trying to formulate rules showed that the complexity especially of having to model several rules at once is very high. Therefore, we devised a structure to the rules and planned to test models

with single or few rules applied at the same time. In the following, we will formalize our procedure. The rules will be formulated in three components:

- An optional condition
- At least one phone to be modified
- The phone it is mapped to — this may also be an empty phone (which would result in a deletion of the phone to be modified)

We choose to visualize this the following way which is inspired by phonological rules[1]:

$$\mathrm{A} \longrightarrow B \ / \ \mathrm{CONDITION}$$

The above formula may be read as phone A is mapped to phone B, if $CONDITION$ applies. Conditions may also be concerned with the spelling of the word, since spellings may carry useful etymological information that is not available from the pronunciation alone. At this point, the reader may correctly interject that reliable etymological information could be gathered from other sources. This option was considered but could not be successfully implemented via the available resources[2]. The information available via the spelling can only be used to check for the existence of graphemes like <Ei ~ ei>, but unfortunately it is not possible to align graphemes to the phones. Following these considerations, three kinds of sub-conditions will be used:

- The phone occurs after X (after ⟦X⟧),
- the phone occurs before Y (before ⟦Y⟧),
- the spelling contains Z (∃<Z>).

Within the CONDITION, X, Y, Z may be a disjunction of items like a set of phones ⟦m ~ n ~ N⟧ (the set of nasal phones known to the AM) or a set of graphemes <Ei ~ ei>. Each sub-condition can be negated (¬). A conjunction of zero to three sub-conditions will be used as the final condition. Consider the following purely illustrative example:

⟦aI⟧ ⟶ ⟦o:6⟧ / after ⟦S ~ s ~ z⟧
∧ (¬ before ⟦I⟧)
∧ ∃ <Ei~ei>

[1] This notation style takes inspiration from the common notation used for phonological rules which is described by Chomsky and Halle (1968, p. 18). It is, however, not equivalent.

[2] A significant reason was that the ASR lexicon for this architecture must list all word forms and word formations separately. For this reason, the percentage of words for which etymological data could be gathered in this work was not sufficient.

Application:

<Seife>	⟦z aI f @⟧	⟶ <Seife> ⟦z o:6 f @⟧
<Design>	⟦d i z aI n⟧	will not be mapped because it lacks <Ei~ei>
<Seil>	⟦z aI l⟧	will not be mapped because ⟦aI⟧ is before ⟦l⟧

If phones are added to the lexicon file for which the acoustic model is not trained, these cannot be processed. This means that the Bavarian sounds must be mapped to the (allo-)phone set of the acoustic model. Consequently, the target phone must be chosen with linguistic consideration of both Standard German and Bavarian. The relationship between the two languages' sound systems is likely distorted in an unknown degree due to factors such as transcription inaccuracy and G2P inaccuracy in domains such as anglicisms that can be seen in Pritzen's work (2021, pp. 6,23–26).

The new pronunciations resulting from a phonetic mapping rule will be added to the baseline lexicon without removing the original pronunciations for multiple reasons. The level of dialectality is a spectrum and Standard German segments may exist in the data. Loanwords also may behave differently than native Bavarian words. This list is not exhaustive but shows that removing Standard German pronunciations may have unpredictable effects which would in turn make our findings less conclusive.

The new lexicon will be created using the original pronunciation and new pronunciations of one or multiple of the rules applied. After initial considerations, the decision was made to focus on testing a small number of rules or individual rules in this explorative experiment. This will limit the percentual effect of the rule-based approach, but it will show if it is feasible and how to implement it. Using the new lexicon and the baseline AM and LM, we will build a hybrid model. This hybrid model will be validated on the validation set (Oachkatzl-dev). If our benchmarking system shows a deterioration of the WER, the model will not be considered for testing. The validation set will also be used to select the most successful target phone(s) if several target phones seem similarly suitable. The validation set is used to possibly reduce the number of models to be evaluated on test sets. This was done to reduce the influence of chance and overfitting on the output by reducing the number of tests on the test set.

Undoubtedly, the phonological history of the German dialects has created a larger phonetic distance between some corresponding sounds in the standard variety and Bavarian than between others (e.g., stdg. *besser*, bav. *bessa* and stdg. *Stein* bav. *Stoa*). It is also necessary to consider that mapping specific differences will have such an unpredictable outcome that these rules are not considered in the scope of this work (like the Bavarian short *o* being as close as the long *o* in Standard German). Therefore, we selected a set of sound differences that are likely to improve the WER and whose effect we assume to be beneficial.

The selection was based on linguistic considerations such as the following (see 2.2 Phonetic Alphabets):

- The articulatory distance of the Standard German and Bavarian phoneme — /a̲/ and /ɔ/ (IPA) differ in backness, roundedness, closeness
- The existence of a likely more suitable phone in our model — ⟦O⟧ for ⟦a⟧ (BasSAMPA for our model's phonetic variable)

The selected candidates will be further investigated in the execution of the experiment. The variables selected for investigation are listed in Table 3.

Table 3 Selected Phonetic Variables

Selected Phones in Lexicon (BasSAMPA)	**Bavarian sounds** (quasi-traditional spellings[1])
⟦a⟧ ⟦a:⟧	Bavarian back rounded vowels, e.g., *å*
⟦aI⟧[2]	Bavarian diphthongs *ei, oa*
⟦aU⟧	Bavarian monophthong *à*
⟦E: ⟧	Bavarian low front *à*
⟦Y⟧ ⟦OY⟧ ⟦9⟧ ⟦2:⟧ ⟦2⟧ (⟦96⟧ ⟦2:6⟧)	Bavarian unrounded front vowels *e, i, ei*
⟦p⟧ ⟦t⟧[3]	Bavarian lenition ('softening') *b, d*
⟦n⟧ ⟦x⟧ ⟦C⟧	Loss of morpheme final consonants
⟦l⟧	l-vocalization *j, i*

Execution

In the following, the execution of the experiment will be elaborated. For more information about the linguistic reasoning behind the mappings, see Sections 2.1 – 2.3. We follow

[1] Consider this spelling as an example: https://www.bayrisches-woerterbuch.de/bairisch-lernen/aussprache-bayrisch/

[2] Mapping the diphthong ⟦aI⟧ is complex due to multiple possible target phones and a particularly large uncertainty how it should be mapped. This also applies to the phones ⟦u:⟧, ⟦i:⟧, ⟦y:⟧ that like ⟦aI⟧ would be mapped to opening diphthongs. Since the phonetic distance is greatest in the case of ⟦aI⟧, only this phone was investigated in this work.

[3] ⟦k⟧ was not investigated since we expected this additional phone to provide little insight.

fundamentals of the linguistical subject such as described by Furhop and Peters (2013), Zehetner (1985) and Niebaum and Macha (2014).

The phones [ɐ̯] and [ɐ̯ː], which are low-central and unrounded vowels in Standard German, largely correspond to Bavarian back rounded vowels. Therefore, we decided to map ⟦a⟧ and ⟦aː⟧ to ⟦O⟧ and ⟦Oː⟧ respectively. We tested these rules separately since we expected a clearly positive impact for the short sound, while the long sound may be influenced by more factors — such as historical differences in vowel shortening and lengthening (cf. stdg. *s*[ɐ̯ː]*gen,* bav. *s*[ɔ]*ng*), negative influences since the long vowel [ɔː] in German is not native but limited to loanwords, and because the long vowel is derived from two separate vowels in Middle High German (*â*, *a*). We expected some negative influences on the rule in positions before nasals and excluded this sound environment from the rules. That is why the final rules were formulated as:

$$⟦a⟧ \longrightarrow ⟦O⟧ \,/\, \neg \text{ before } ⟦m \sim n \sim N⟧,$$

$$⟦aː⟧ \longrightarrow ⟦Oː⟧ \,/\, \neg \text{ before } ⟦m \sim n \sim N⟧$$

Mapping ⟦aI⟧ proved to be more complex for several reasons. There are two possible correspondences for this diphthong in Bavarian. These are not predictable from Standard German anymore since among these two only Bavarian preserves the distinction. The Standard German orthography also does not reflect the difference (e.g., mhg. *wîz* > *weiß*[1], bav. *weiß,* mhg. *weiz* > *weiß*, bav. *woaß*). A different sound change leads to a substantial phonetic difference between Standard German and Bavarian: The common realization *oa* [oɐ̯̃] has vastly different features from [ɐ̯ɪ̯̃] apart from them being diphthongs. This shows the need to model this difference in the lexicon. The other correspondence of [ɐ̯ɪ̯̃] is much closer to the Standard German pronunciation. We assume that it is unlikely that we will find a more suitable mapping for the latter, so it will not be mapped to another phone sequence in this work.

The focus instead lies on the former diphthong, here transcribed as [oɐ̯̃]. Regarding the distribution of the two diphthongs, we can exclude a great number of false predictions by limiting the rule to words spelled with a native spelling of the diphthong, such as <ei ~ ai>. This will exclude many loanwords such as anglicisms from the application of the rule (e.g., *online*). Another challenge is that in word formation via compounding, the diphthong [ɐ̯ɪ̯̃] can occur many times in a single word. As stated before, the spelling cannot be aligned to the phones. Predicting which of the two Bavarian diphthongs occurs is not possible in most cases. This means that for words that have n instances of ⟦aI⟧, the number of word forms needed to always model the correct Bavarian form is 2^n. From these, all but one would be incorrect, which does not seem to be in the interest of this experiment. We therefore chose to apply the

[1] This notation may be read as *Middle High German wîz became weiß.*

rule to all instances of ⟦aI⟧. We argue that this procedure is ideal for this experiment. It can not only improve the modeling for Bavarian words where it would be correct to map all n instances of ⟦aI⟧ this way, but it is reasonable to assume that it improves the modeling for all words where this applies to at least half of the instances of ⟦aI⟧. The procedure may not be optimal, which should be investigated in future work.

Finally, the selection of a well-suited representation of the diphthong will require several options to be validated, because only this way can we be sure that the mapping rule is appropriate. We therefore formulated four rules for validation. ⟦o:6⟧ is commonly pronounced[1] as an opening diphthong in Standard German, resulting from a vocalization of the phoneme /r/ (e.g., *Ohr* [oːɐ̯], but /oːr/). Similarly ⟦O6⟧ was chosen. As options without phonologically underlying[2] /r/, ⟦o: a⟧ and ⟦O a⟧ were chosen. The advantages and disadvantages of these phones cannot be discussed in the scope of this work, but the validation process ensures that a suitable representation is chosen. The final rules are listed below:

⟦aI⟧ ⟶ ⟦o:6⟧ / ∃ <Ei ~ ei ~ Ai ~ ai>,

⟦aI⟧ ⟶ ⟦O6⟧ / ∃ <Ei ~ ei ~ Ai ~ ai>,

⟦aI⟧ ⟶ ⟦o: a⟧ / ∃ <Ei ~ ei ~ Ai ~ ai>,

⟦aI⟧ ⟶ ⟦O a⟧ / ∃ <Ei ~ ei ~ Ai ~ ai>

For ⟦aU⟧, the only mapping target is the ⟦a:⟧. The condition of <Au ~ au> was added to avoid mapping English loanwords.

⟦aU⟧ ⟶ ⟦a:⟧ / ∃ <Au ~ au>

The mid-low [ɛː] often corresponds to a much lower front vowel in Bavarian. We can only map it to the mid [a̝ː]. This is less likely to decrease the phonetic distance than in the previous rule due to the loss of a feature that both the Bavarian and Standard German phone share. They both are front vowels, but we must map the Bavarian open front vowel to an open central vowel. Since this is a realization of a German umlaut, we can limit the rule to words with <Ä ~ ä>:

⟦E:⟧ ⟶ ⟦a:⟧ / ∃ <Ä ~ ä>

[1] This assumes that [o:6] is a phonetic variable of the AM and is not strictly equal to IPA [oːɐ̯]. We do not know what it precisely stands for, but that, due to its training, it must strongly correlate with the pronunciation [oːɐ̯].

[2] In this case, we aimed to choose a neutral wording. We do not contest different analyses regarding morpho-phonology, phonology, structural phonology and generative phonology.

The German umlauts represented by <ö> and <ü>, as well as the diphthong ⟦OY⟧ went through unrounding processes in Bavarian, which caused a merger with the unrounded front vowels. The following rules were created.

$$⟦Y⟧ \longrightarrow ⟦I⟧ \,/\, \exists \ \langle Ü \sim ü \rangle,$$

$$⟦OY⟧ \longrightarrow ⟦aI⟧ \,/\, \exists \ \langle Äu \sim äu \sim Eu \sim eu \rangle,$$

$$⟦9⟧ \longrightarrow ⟦E⟧,\ ⟦2⟧ \longrightarrow ⟦e⟧,\ ⟦2:⟧ \longrightarrow ⟦e:⟧,\ ⟦96⟧ \longrightarrow ⟦E6⟧,\ ⟦2:6⟧ \longrightarrow ⟦e:6⟧$$

The latter five do not have a condition, which is not ideal since French loanwords also may have these phones, but this cannot be changed at this point. The first two rules were applied to create individual models, the final five were used to create a single model. This means that only unmodified words and words modified with all applicable rules will be in the lexicon. These differences may be considered an inconsistency to the other models that is due it being the earliest of the models.

Regarding consonantal rules, the phones ⟦p⟧ and ⟦t⟧ were mapped to possibly better model the Bavarian realization of these consonants. While the Standard German opposition between the consonant set /p, t/ and the set /b, d/ is characterized by a set of features such as aspiration and voicing (Furhop & Peters, 2013, p. 42), the Bavarian correspondences of /b, d/ are described to be always unvoiced and the distinction between the pairs /p/ - /b/ and /t/ - /d/ to be (largely) lost (Zehetner, 1985). To model this, we formulated rules to map the phonetic entries ⟦p⟧ and ⟦t⟧ to ⟦b⟧ and ⟦d⟧ and limited the rules to an environment before vowels. This way, we aim to avoid negative results of the word-final loss of the distinction in Standard German that can be called *final obstruent devoicing* (stdg. *Auslautverhärtung* 'final fortition') (Furhop & Peters, 2013, p. 73; Brockhaus, 2012, p. 3). This can model some of the environments mentioned by Zehetner with the notable exception of /pr/ and /tr/ (1985, pp. 55,84). Another negative influence we foresaw was that after sibilants (s-like consonants, here ⟦s ~ S⟧), the plosives in this series are unvoiced but unaspirated and there is also no opposition in the pairs /p/ - /b/ and /t/ - /d/. We therefore excluded the consonants from mapping that were preceded by these phones — e.g., ***Sp****iel* ⟦S p⟧*,* ***St****adt* ⟦S t⟧*,* ***St****il* ⟦s t ~ S t⟧. Due to the high number of vowels, we will not list them all but define the variable V as such.

$$⟦p⟧ \longrightarrow ⟦b⟧ \,/\, (\neg \text{ after } ⟦S \sim s⟧) \ \wedge \text{ before } V$$

$$⟦t⟧ \longrightarrow ⟦d⟧ \,/\, (\neg \text{ after } ⟦S \sim s⟧) \ \wedge \text{ before } V$$

Central Bavarian, such as the dialect of Munich, is characterized by a vocalization ('becoming a vowel') of the historical phoneme /l/ in final position. This dialect feature most likely leads to a significant modeling problem for the ASR system due to two factors: Not only does the stable consonant phoneme /l/ of Standard German correspond to vowels and

diphthongs in Bavarian (stdg. *halt* [hal̪tʰ]- bav. *hoit* [hɔi̯t̪]), but it can also modify the preceding vowel (*viel* [fiːl] – *vui* [fui̯]). The significance of this consonantal rule and the existence of certainly more similar mapping targets in the phonetic lexicon (⟦j⟧, ⟦i⟧, ⟦aɪ⟧, ⟦OY⟧) made us select this variable for further investigation. Nevertheless, the creation of the rule was not successful in the scope of this work. The extreme complexity of this rule might be offset by its utility, but this does not account for another disadvantage: The rule is entirely dependent on the syllable structure of the word, since the /l/ in the syllable onset was not affected by the sound change, whilst final /l/ was. There is, however, no consistent information on the syllable borders in the phonetic lexicon. Complex approximations might be made to account for the syllable borders or the rules could be limited to the final phone in the word. But due to the challenges overall, we considered this variable infeasible for the scope of this work and the lack of syllable information in the source data as being critical.

Similar considerations apply to the final loss of phones such as ⟦x⟧, ⟦C⟧, ⟦n⟧. These are also reliant on syllable information, although modeling them word-finally may be advantageous on its own. However, we found that removing these phones alters the syllable structure of the words itself. This would result in illegal syllables in German such as ⟦Q I⟧ because these vowels only occur in closed syllables (ending in a consonant). In this work, we limited the complexity of the rules to modifying a single phonetic variable (including diphthongs) at a time. For this reason, the final two rule variables were left to future work.

Observation

Table 4 Benchmark Results of Rule-Based Lexicon Additions

Variable	**Target Phone**	**Substi-tutions**	**Oachkatzl-test WER (%)** (5.32h, 50,403W)		**Bairisch-2020-Altb WER (%)** (0.15h, 1,672W)		**DSA-Wenker WER (%)** (0.22h, 2,482W)	
BASELINE			52.96	**Rel.**	39.59	**Rel.**	57.25	**Rel.**
⟦a⟧	⟦O⟧	1.1m	52.49	**-0.89**	38.82	**-1.94**	57.05	-0.35
⟦a:⟧	⟦O:⟧	585k	52.93	-0.06	39.53	-0.15	57.25	0.00
⟦aI⟧	⟦o: a⟧	469k	52.72	-0.45	39.47	-0,30	56.85	-0.70
⟦Y⟧	⟦I⟧	103k	52.97	+0.02	39.77	+0.45	57.17	-0.14
⟦9⟧ ⟦2⟧ ⟦2:⟧ ⟦96⟧ ⟦2:6⟧	⟦E⟧ ⟦e⟧ ⟦e:⟧ ⟦E6⟧ ⟦e:6⟧	98k	52.94	-0.04	39.35	-0.61	56.77	-0.84
⟦OY⟧	⟦aI⟧	102k	52.84	-0.23	39.47	-0.30	56.08	**-2.04**
⟦aU⟧	⟦a:⟧	236k	52.93	-0.06	39.35	-0.61	57.21	-0.07

For two of the three planned rule sets of consonantal mapping, significant interdependencies were found during the experiment. Therefore, no models could be made for them in the scope of this work. The rules for ⟦E:⟧, ⟦p⟧ and ⟦b⟧ failed the validation process by increasing the WER on the validation set. During the validation of the mapping rules for ⟦aI⟧, ⟦o: a⟧ was the best-performing target phone sequence. The achieved WERs on the test sets are shown in Table 4.

To provide some more statistics for this approach: For ⟦a⟧, approximately 1,100,000 pronunciations were generated, resulting in 237 more successfully recognized words on Oachkatzl-test. For each model based on the diphthong ⟦aI⟧, approximately 470,000 pronunciations were generated, resulting in 121 more recognized words using the successfully validated model.

Discussion of the Findings

We observe the vowel mappings to be feasible and largely effective. One vowel mapping rule did not pass the validation phase and the mapping of one vowel, ⟦Y⟧, had a negative impact on two test sets.

The investigated consonant mappings are either not beneficial or infeasible in the scope of this work due to complex interdependencies and the lack of information on syllable borders.

Modeling the opening diphthong *oa* [oa̯] was surprisingly most successful using two phones ⟦o: a⟧ instead of the phones ⟦o6⟧ or ⟦O6⟧, by a small margin.

Applying these individual rules or a small set of rules has a small effect on the WER, which may strictly be considered statistically insignificant. To verify this, a simple significance test with the significance level $\alpha = 0.05$ was conducted for each model-dataset pair individually. As expected from such low improvements, considering the test set sizes and a baseline WER close to 50%, the significance could not be supported with a probability of 95%. In fact, the change in WER due to the mapping of ⟦a⟧ could be considered significant with only a probability of 90% on Oachkatzl-test. Nevertheless, we argue that a priori only a small subset of the (allo-)phone inventory of the AM is known to be altered. Therefore, it may be assumed that the improvement regarding the recognition of the mapped phones as well as the words that can be affected by the mapping is higher. Hence, the significance would likely be more certain when only observing the words that could be influenced by the change. In a similar condition, Pritzen used an entity error rate for anglicisms (which were the only altered lexicon entries), allowing for a more precise observation of the changes (2021, p. 32).

At the onset of this approach, finding a way to perform phonetic mapping proved much more challenging than initially assumed. This is also the reason why the process took significantly longer initially. We interpret it as a success of this work that a workflow and a formalization of this process could be achieved. Combined with the creation of a Bavarian validation set, individual rules can be selected before they are further used in testing or for combination with other rules. The results of the vocalic mapping rules show improvements and new problems with the consonantal rules could be identified and better handled.

The successful vowel mappings are candidates for a combined rule-based model. It is most likely that such a model would show a performance change that is not merely a sum of the currently seen improvements. The performance may be even higher since there is a threshold for the improvement before a new word is output and the WER improves. This threshold may not be reached by one phone mapping alone, but it is likely when many are combined. On the other hand, a smaller improvement might be observed if additional mappings have reduced returns.

Due to rule-interdependencies, we could only anticipate a suitable combination process after successful validation, at which point it was not possible to combine the rules and benchmark them for inclusion in this work anymore. This must be left to future work. One challenge of combining rules is that one rule may influence another. This is most challenging when there are rules that involve more than one phone due to the CONDITION, as we called

it, or the phone to be mapped itself (e.g., ⟦h **aI** s t⟧ ⟶ ⟦h **o: a** s t⟧, but not ⟶ ⟦h o: **O** s t⟧). Even if only one phone is affected, every segment of a word may only be mapped once (e.g., ⟦h **OY** z 6⟧ ⟶ ⟦h **aI** z 6⟧, but not ⟶ ⟦h **o: a** z 6⟧) or the ideal order of the applied rules must be determined, which is difficult to assess. Therefore, creating a model with all rules combined must be left to future work.

4.2 Data-Driven Approaches

Low-resource environments are at a major disadvantage when using data-driven approaches to train ASR. Training a Bavarian model from scratch may be called futile if less than thirty hours of transcribed training data are available. However, many different approaches loosely grouped under the term transfer learning (Wang and Zheng 2015) have shown that it is possible to transfer knowledge from previous training on different tasks to a similar task with little training data. This way, a low-resource task can be learned more successfully. In the case of Bavarian ASR bootstrapped from a Standard German model, we can benefit from a particular similarity between the two languages. Therefore, data-driven approaches to Bavarian ASR will be explored in this chapter.

4.2.1 Training of a Hybrid Model

We will begin with data-driven modifications of the hybrid baseline model. For this, both training the acoustic model and the language model are evaluated. The results for these training experiments can be found at the end of this subsection. Four setups will be compared: the Baseline-ASR, AM-Training, LM-Training, AM&LM-Training.

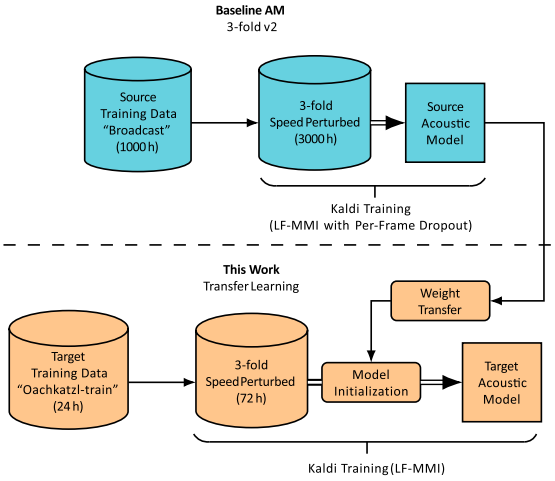
4.2.1a Acoustic Model

Gref (2022, p. 133) and Gref et al. (2019) investigated a transfer learning approach for the Fraunhofer IAIS AM which showed potential improvements even with little training data, such as 3.5 hours of challenging audio data (2022, p. 58). The AM of the baseline is called 3-fold v2; more information can be found in Chapter 2 Fundamentals, and in Gref's dissertation (2022).

Experimental Design:

Figure 8 is ef et al. (2019, Figure 1) and displays the transfer learning in this experiment.

Figure 8 Hybrid TL Procedure (inspired by Gref et al., 2019)



We perform transfer learning on the baseline AM "3fold v2". The procedure corresponds to stage 2 in Gref's research (2019). We conduct transfer learning via weight transfer, keeping the output layer of the original model. Training is conducted using the Kaldi toolkit for four epochs, with an initial learning rate of $3 \cdot 10^{-6}$ and a final learning rate of $3 \cdot 10^{-7}$. No dropout is applied due to the small size of the training set (ibidem). This transfer learning process has been extensively researched (Gref et al., 2019; Gref, 2022, p. 133); the hyperparameters will, therefore, not be altered in this work.

The source system "3fold v2" is an AM trained on broadcast data (see 2.6.1 Standard German Baseline Models), which is a significantly different domain and variety of German than the target Bavarian dialect data. The AM will be trained on Oachkatzl-train using the LF-MMI criterion (see 2.6.1 Standard German Baseline Models). This training set is several times larger than the 3.5 hours of oral history interviews used in TL experiments by Gref and Gref et al. (2019; 2022, p. 136), which is promising. It remains to be seen if the knowledge can be transferred efficiently to the much different phonological structure. For the training procedure, the model is supplied with an automatically generated lexicon. The transcriptions are generated by a Fraunhofer-IAIS-internal G2P model trained with Sequitur G2P (see 2.6.1 Standard German Baseline Models). The trained model will finally be evaluated on the three test sets.

4.2.1b Language Model

As we elaborated in Section 1.2 Problem Definition, there are differences between Bavarian and Standard German that are for a large part or completely modeled by the LM. It follows that in the case of our hybrid model, the ASR performance on Bavarian is limited in part due to the inaccurate modeling of the n-grams. This will be investigated here.

Experimental Design

The modeling performance of the baseline language model must initially be determined using suitable metrics. For this, the perplexity and running OOV rate will be determined on the test sets. It was considered to calculate this on the validation set, however, this would not provide information on two of the different domains represented by the three test sets.

A priori, we hoped to reach a perplexity of 700-1,000 in this challenging domain. While smaller values would be better, we cannot expect this for such dialectal data. Values between 700 and 1,000 would be comparable to the Oral History data in Gref's work and, therefore, may be considered acceptable, especially compared to the "Challenging Broadcast" datasets (2022). Still, we must mention that Garner et al. considered a perplexity of 613 (in their case on a test set of 463 sentences) high (2014). Table 5 shows the perplexities for two LMs, with "Default LM" being a 500,000-word LM and "Large LM" a 2-million-word LM (Gref 2022, p. 72).

Table 5 Perplexity of the LMs in Gref's Work (adapted from Table 3.8, 2022, p. 73)

Dataset	**Default LM**	**Large LM**
DiSCo Planned Clean	715.13	690.75
DiSCo Planned Mix	847.76	728.54
DiSCo Spontaneous Clean	494.24	576.27
DiSCo Spontaneous Mix	575.20	628.23
German Broadcast 2016	639.68	524.23
Challenging Broadcast	1,152.90	946.13
Oral History	769.56	767.27
Interaction (Linguistics)	1,480.74	1,748.00
Spoken QALD-7	1,626.62	1,097.35

Similarly to Gref (2022, p. 75) we will judge the baseline LM and the trained LM according to their running OOV rate in relation to values of normal in-domain data, challenging in-domain data ("Challenging Broadcast") and out-of-domain data provided by Gref (ibid.). Due to the large gap in the running OOV rates between "Challenging Broadcast" and the other sets, we will consider values higher than 2.27% problematic.

Table 6 Running OOV Rates of two LMs across Test Sets (adapted from Table 3.9, Gref 2022, p. 74)

Language Model	Running OOV Rate	
	Default	**Large**
DiSCo Planned Clean	0.75%	0.30%
DiSCo Planned Mix	1.20%	0.46%
DiSCo Spontaneous Clean	0.44%	0.21%
DiSCo Spontaneous Mix	0.70%	0.22%
German Broadcast 2016	1.39%	0.45%
Challenging Broadcast	2.27%	0.82%
Oral History	1.72%	0.96%
Interaction (Linguistics)	1.46%	0.72%
Spoken QALD-7	1.57%	0.68%

Table 7 Baseline LM Perplexity and Running OOV Rate

Measure	**Oachkatzl-test** (5.32h, 50,403W)	**Bairisch-2020-Altb** (0.15h, 1,672W)	**DSA-Wenker** (0.22h, 2,482W)
Perplexity	1,708.05	945.78	1,009.71
Running OOV Rate	0.85%	1.26%	0.28%

The baseline performance regarding perplexity and OOV rate can be seen in Table 7. It follows that the baseline running OOV rate is not problematic. The small Bairisch-2020-Altb test set is modeled acceptably for a challenging out-of-domain test set, judging from its perplexity below 1,000, but not well. However, the large test set Oachkatzl-test has a perplexity we must consider too high. Compared to all configurations in Gref's work (2022, p. 73) and even challenging out-of-domain datasets, it is the second highest perplexity value. We conclude that language model training is an effective means to improve the speech recognition for these samples.

Since the target output of the ASR system is orthographic Standard German, Bavarian text is not suitable for training the language model. Instead, Standard German transcription data for Bavarian, as it is used in the training set, is appropriate. Therefore, the whole data from Oachkatzl-train will be used for training the language model. Moreover, we are not limited by requirements of speech-to-text training sets because no audio is necessary for

training. Therefore, even misaligned and inaccurate subtitle data may be beneficial. For this reason, we will use the training set Bavarian-ASR-train for LM training. While it could not be cleaned to a level that it could fulfill its previously intended role, this does not matter here. As can be seen in Section 3.3 Overview of Bavarian Datasets, the dataset has 629,323 words and an average of 8.6 words per segment. The number of words is small considering that it amounts to less than 1% of the Default LM seen in Gref's work (2022, p. 72). Therefore, it was considered to use LM interpolation to combine the trained model with the baseline LM model while giving the smaller model a higher weight. Due to time constraints, the LM interpolation must be left to future work. We hope that the experiment, as it is conducted, still provides useful insights to the reader in the context of this work.

As for the baseline-LM (cf. Gref 2022, p. 71), we will use the IRSTLM-toolkit (Federico, Bertoldi, & Cettolo, 2008) to train a 5-gram model, also applying the pruning threshold of $1 \cdot 10^{-8}$ of the baseline model. For the actual training data, we will concatenate the training set ten times, due to the small size of the training dataset — following a personal recommendation from a colleague. After training, a new model will be built using the baseline AM and lexicon and the new LM. It will be evaluated using three measures: the WER, the new perplexity and the running OOV rate, determined on the three test sets. We include the linguistic test set "DSA-Wenker". However, due to the sentences being created for the extraction of dialect features, the results seem less meaningful than in the case of acoustic modeling, where the task is to recognize a small set of speech sounds.

Regardless of the results of the WER and perplexity, we must expect an increase of the OOV rates since the training data is so small.

4.2.1c Benchmarking

Three hybrid ASR system were created with the trained models in order to evaluate the training. The configurations are listed in Table 8.

Table 8 Hybrid ASR System Configurations

Experiment	**AM**	**LM**
AM Training	This work	Baseline
LM Training	Baseline	This work
AM+LM Training	This work	This work

Observation

Table 9 shows the benchmarking results of the baseline model and the models created in this chapter:

Table 9 Benchmarking Results Hybrid Training

Variable	**Oachkatzl-test WER** (5.32h, 50,403W)		**Bairisch-2020-Altb WER** (0.15h, 1,672W)		**DSA-Wenker WER** (0.22h, 2,482W)	
Baseline	52.96%	**Relative**	39.59%	**Relative**	57.25%	**Relative**
AM Training	43.36%	-18.13%	33.19%	**-16.17%**	46.05%	**-19.56%**
LM Training	46.74%	-11.74%	47.07%	+18.89%	57.86%	+1.07%
AM+LM Training	39.33%	**-25.74%**	42.28%	+6.79%	49.56%	-13.43%

Table 10 lists the perplexity and running OOV rate for the baseline model and the models created in this chapter.

Table 10 Trained LM Evaluation Results

Test Set	**Baseline PP**	**Bav.-LM PP**	**Change**	**Baseline OOV**	**Bav.-LM OOV**	**Change**
Oachkatzl-test (5.32h, 50,403W)	1,708.05	**492.57**	-71%	**0.85%**	3.13%	+73%
Bairisch-2020-Altb (0.15h, 1,672W)	**945.78**	1,624.85	+42%	**1.26%**	8.24%	+85%
DSA-Wenker (0.22h, 2,482W)	**1,009.71**	1,094.29	+8%	**0.28%**	3.22%	+91%

As seen in Table 9, the AM training, as conducted in this work, had the effect of reducing the WER across all test sets. The results show the lowest improvement on Bairisch-2020-Altb with 16,17%. Bairisch-2020-Altb also shows both the best WER in this experiment and the best WER using the baseline model.

The Bavarian LM could reach a WER improvement of 11.74% on the Oachkatzl-test set. On the other test sets it has a negative influence compared to the baseline. The influence was 1.07% for the variable LM training and 18.89% on Bairisch-2020-Altb.

The ASR system using both a trained AM and LM reached the highest relative WER improvement in this work (-25.74%) and bringing the ASR system just below the 40% WER

mark. The effect of the AM is seen to be positive if the LM training is beneficial on its own. If the LM training is detrimental, the performance of the the model AM+LM Training is worse than of the model AM Training in this work.

The baseline LM shows a high perplexity of 1,708.05 on Oachkatzl-test whereas the new LM shows a much lower perplexity of 492.57, which corresponds to an improvement of 71%. The ASR benchmark results deteriorated by 42% on Bairisch-2020-Altb and by 8% on DSA-Wenker.

Interpretation

The results show that using only the Bavarian-ASR-train data is tremendously effective for Oachkatzl data (likely also for all "Dahoam is Dahoam" data, which was not tested here), but causes worse modeling for the small test set Bairisch-2020-Altb. One of the reasons for the different effects may be the different domains. More importantly, the running OOV rate for the test set Bairisch-2020-Altb reached a value of 8.24%. This means that about every twelfth encountered word in the small test set is unknown. We must deem such high values unacceptable and see this as a probable reason for the high perplexity increase. To interpret the reason for the difference between the results on different datasets, we assume that while the other two test sets are at least partly scripted, Bairisch-2020-Altb can be considered a sample of a quite special domain. Also, the effects are probably increased by the small size of the training corpus, and are likely statistically increased by the small size of Bairisch-2020-Altb. The 100 segments are from an interview situation regarding the interviewee's personal experiences and opinions in a niche sports activity. Specialized vocabulary like *Supraspinatussehne* ('supraspinatus tendon'[1]) can explain the inadequacy of a small language model for this dataset as such words are unlikely to occur in the training data. We also must consider that Bairisch-2020-Altb is shown to be modeled best by the baseline ASR system in two significant measures: the WER and the perplexity. It is the only test set on which the baseline system achieves a WER below the 40% margin we aimed to achieve in this work. It is also the test set on which the best WER score in this experiment is achieved (33.39% WER). It may be assumed that an improvement using a new LM is less likely, and a performance degradation is more likely.

It is clear that the small language model lacks robustness. As initially intended, an interpolated LM would need to be created. This result was expected since it turned out that the limited language model would need to be used in this work. Despite this significant shortcoming, the improvement of the perplexity on the largest test set Oachkatzl-train was more significant than expected given the training data. This may be caused by the nearly identical domain, since most segments from both the training data and Oachkatzl-test seem

[1] https://www.dict.cc/?s=supraspinatussehne

to be taken from the series “Dahoam is Dahoam”. To be sure that the clear advantage this data has is not due to segments shared across the training and testing data, we consulted the institute’s Benchmark Viewer and manually compared some sentences that had their WER substantially improved by the new LM. We found that many segments had improvements of multiple consecutive words, some of which raised the suspicion that these may be caused by shared 5-grams or even whole segments in training data and testing data. However, the n-grams could not be found in the test data. This provides the assurance needed that the trained n-gram model can predict unknown segments in the testing data well. If future research aims to create a workable LM instead of a prototype, this should be statistically verified.

We conclude our discussion of the LM by acknowledging that the model in its current state is only a prototype and not useful in practice, but that the improvements achieved on Oachkatzl-test are greater than expected and promising for the development of a specialized ASR system for data like “Dahoam is Dahoam”. The results indicate that the selected data for LM training can help recognize challenging data. Increasing the amount of German data with broadcast data or other sources while giving the Bavarian training data more weight could result in a more robust LM to predict Bavarian samples better.

The results for the trained AM show that transfer learning using approximately 24 hours of Bavarian data is a viable approach to improving Bavarian ASR. On all data, AM training shows substantial improvements. We can infer that using a Standard German lexicon, the complex phone correspondences between Standard German and Bavarian can be learned to a degree so as to likely improve the WER by a relative 15-20%.

4.2.2 End-to-end Fine-Tuning using Wav2vec 2.0

Via its novel pre-training objective, wav2vec 2.0 has shown improvements to the end-to-end paradigm, especially in the low-resource domain (Baevski et al., 2020). Due to their promising results (ibid.), wav2vec 2.0 seems equally suitable for the task at hand as our hybrid approaches. We will therefore explore if fine-tuning a wav2vec 2.0 model with our training data is a feasible means of improving Bavarian ASR.

Experimental Design

The model internal-IAIS-wav2vec2 is a model fine-tuned for German and it was created by applying further training to the model grosman2021-wav2vec2 (see 2.6.1 Standard German Baseline Models). Most notably, internal-IAIS-wav2vec2 was trained on the same large broadcasting data as the Fraunhofer IAIS hybrid system: Ger-TV-1000h. For comparison, the public source model grosman2021-wav2vec2 will be benchmarked and fine-tuned. As a first assessment, it seems relevant to compare the performance of both wav2vec 2.0 models to the baseline hybrid model on the Bavarian test sets. Comparing the Bavarian ASR performance

of these models trained for Standard German will provide crucial information on the differences of the models and if the results will be comparable.

Table 11 Comparative Bavarian ASR performance of German Hybrid and Wav2Vec 2.0 Models

Model	**Oachkatzl-test WER** (5.32h, 50,403W)		**Bairisch-2020-Altb WER** (0.15h, 1,672W)		**DSA-Wenker WER** (0.22h, 2,482W)	
Hybrid Baseline	52.96%	**Relative**	39.59%	**Relative**	57.25%	**Relative**
Grosman2021-wav2vec2	66.28%	+25.15%	60.64%	+53.17%	71.31%	+24.56%
Internal-IAIS-wav2vec2	49.76%	**-6.04%**	36.00%	**-9.07%**	56.00%	**-2.18%**

Table 11 shows the comparative performance of the three baseline models in this work. Naturally, the relative performance is not a change in performance contrary to most overviews in this work. It is notable that the model grosman2021-wav2vec2 invariably performs worse on these test sets than the hybrid baseline by about 25% to more than 50%. On the other hand, internal-IAIS-wav2vec2 invariably performs better than the hybrid baseline, yet by a much smaller margin. Using a simple significance test on the individual test pairs, we can conclude that all configurations except the performance difference between the hybrid baseline and internal-IAIS-wav2vec2 on DSA-Wenker are significant with $\alpha = 0.05$. It is therefore possible to conclude that without further training, grosman2021-wav2vec2 is significantly inferior to the hybrid baseline in this domain, while internal-IAIS-wav2vec2 is significantly better at modeling our Bavarian test sets, but by a relatively small margin, making it more comparable[1] to the hybrid baseline. For the sake of this work, we estimate the difference of internal-IAIS-wav2vec2 to the hybrid baseline at:

$$\frac{(-6.04\%) \cdot 50403\text{W} + (-9.07\%) \cdot 1672\text{W} + (-2.18\%) \cdot 2482\text{W}}{50403\text{W} + 1672\text{W} + 2482\text{W}} = \frac{-325009.92\%}{54557}$$

$$= -5.97\%$$

In our experiment, we will explore the effect of using our Bavarian training data to fine-tune wav2vec 2.0 models previously fine-tuned on German broadcast data. Due to time-constraints, we will not vary the training hyperparameters other than the number of epochs in this work, but will only use settings previously used in projects at Fraunhofer IAIS.

[1] We will consider the two models comparable in the context of this work.

Using a training batch size of 2, it is possible to train the model on a single GPU with 8GB of video RAM using CUDA[1]. A learning rate of $3 \cdot 10^{-4}$ will be used and as well as 500 (linear) warmup steps. Because of the small size of the training data, no dropout or weight decay will be used (cf. Gref et al., 2019). We will use the default Transformers linear learning rate scheduler from the Hugging Face Transformers library (Wolf et al., 2020). Using the validation set Oachkatzl-dev, six epochs of training on the Oachkatzl-train set will be evaluated for internal-IAIS-wav2vec2 and grosman2021-wav2vec2. The best epoch will be determined. Finally, the corresponding models will be evaluated on all test sets.

For our experiment, we converted all datasets in the Kaldi format into a simple CSV format. While it was advantageous for the Fraunhofer IAIS hybrid model benchmarking workflow to have multiple segments in each audio file, the used scripts for training and inference using wav2vec 2.0 models required one segment per audio file. Another difference is that the wav2vec 2.0 models in this work do not use the character <ß>. Therefore, it was substituted with <ss> in the transcriptions. We must mention that this may have a small effect on the comparability of the wav2vec 2.0 and hybrid models. However, hybrid models do not map the input sequence to single graphemes but to word units taken from the lexicon. Hence, we argue that this would only cause a practical difference between the wav2vec 2.0 and hybrid models in the rare case that a word is distinguished only by the opposition <ss> - <ß>, as in *Massen*, *Maßen*.

Observation

We observed that the first epoch brought improvements to the performance while for epochs 2-6 the performance shrank (even during epoch-internal evaluation). The models trained for one epoch were benchmarked and the results can be seen in Table 12.

[1] https://developer.nvidia.com/cuda-toolkit

Table 12 Benchmarking Results Wav2Vec 2.0, 1 Epoch

Model	**Oachkatzl-test WER** (5.32h, 50,403W)			**Bairisch-2020-Altb WER** (0.15h, 1,672W)			**DSA-Wenker WER** (0.22h, 2,482W)		
	absolute	%-Diff. hybrid	%-Diff. before	absolute	%-Diff. hybrid	%-Diff. before	absolute	%-Diff. hybrid	%-Diff. before
Hybrid base	52.96%			39.59%			57.25%		
Gros. base	66.28%	+25.15		60.64%	+53.17		71.31%	+24.56	
IAIS base	49.76%	-6.04		**36.00%**	-9.07		56.00%	-2.18	
Gros. 1 ep.	42.39%	-19.96	-36.04	45.75%	+15.56	-24.55	49.76%	-13.08	-30.22
IAIS 1 ep.	**36.91%**	-30.31	-25.82	42.11%	+6.37	+16.97	**45.45%**	-20.61	-18.84

Interpretation

Training the models for one epoch on the training set Oachkatzl-train can be viewed as a success since in most cases this led to substantial WER improvements. Particularly high improvements to the baseline model were seen on grosman2021-wav2vec2. This may be explainable by the much lower performance of the model before fine-tuning. In the case of internal-iais-wav2vec2 we saw a WER increase on Bairisch-2020-Altb. We consider this comparable to the findings in Section 4.2.1c Benchmarking where this test set was negatively influenced by our language model training. Since in the end-to-end usage of wav2vec 2.0 the language modeling takes place in the same model, internal-iais-wav2vec2's grammatical modeling was altered by the training. The results in Section 4.2.1c Benchmarking clearly suggest a benefit of the learned acoustic representations. Hence, the non-acoustic language modeling is likely the cause of the deterioration. Even though this explanation seems likely, it may also be considered that the model performance deteriorated on this test set partly due to acoustic overfitting (or loss of Standard German competence).

Since the performance after one epoch of training improved, but deteriorated for the later epochs, it seems likely that the hyperparameters are not ideal for this task. Especially the learning rate, if set too high, would explain why such promising results were achieved in the first epoch, but deteriorated in successive training.

We can conclude that wav2vec 2.0 fine-tuning is a particularly suitable approach to Bavarian ASR. Both the baseline internal-IAIS-wav2vec2 performed significantly better than the hybrid baseline at about -6% WER difference. Fine-tuning this model created the most successful recognizer for the large test set Oachkatzl-test in this work. Fine-tuning gros-

man2021-wav2vec2 using Oachkatzl-train surprisingly also led to a better performing model than the hybrid baseline on two out of three test sets. We see a likely drawback of the wav2vec 2.0 models in the language modeling for test sets such as Bairisch-2020-Altb.

Table 13 Benchmark results for all data-driven approaches

Model	**Oachkatzl-test WER (5.32h, 50,403W)**		**Bairisch-2020-Altb WER (0.15h, 1,672W)**		**DSA-Wenker WER (0.22h, 2,482W)**	
	Absolute	**%-Diff. hybrid baseline**	**Absolute**	**%-Diff. hybrid baseline**	**Absolute**	**%-Diff. hybrid baseline**
Hybrid baseline	52.96%		39.59%		57.25%	
AM Training	43.36%	-18.13	**33.19%**	**-16.17**	46.05%	-19.56
LM Training	46.74%	-11.74	47.07%	+18.89	57.86%	+1.07
AM+LM Training	39.33%	-25.74	42.28%	+6.79	49.56%	-13.43
Grosman2021-wav2vec2	66.28%	+25.15	60.64%	+53.17	71.31%	+24.56
Internal-IAIS-wav2vec2	49.76%	-6.04	36.00%	-9.07	56.00%	-2.18
Grosman fine-tuned	42.39%	-19.96	45.75%	+15.56	49.76%	-13.08
Internal-IAIS fine-tuned	**36.91%**	**-30.31**	42.11%	+6.37	**45.45%**	**-20.61**

Table 13 lists the benchmark results for all data-driven approaches in this work. It must be noted that the difference in performance is given relative to the hybrid baseline model — even for the end-to-end models. Hence, not all the differences are attributable to this work. We compare all models to the same baseline for increased intelligibility of the table and since the hybrid baseline model is the de-facto baseline model for German speech recognition at Fraunhofer IAIS. We consider this comparison especially useful since both the Fraunhofer IAIS hybrid and wav2vec 2.0 models show similar tendencies in their performance (see 4.2.2 End-to-end Fine-Tuning using Wav2vec 2.0).

Finally, Table 14 shows an example of a segment and the predictions of all end-to-end models in this work to help visualize the differences between the models. The segment shows similar challenges across the different models. In the following, we will analyze three examples. There are some differences in capitalization between the models since the used wav2vec 2.0 models do not distinguish casing. However, as mentioned in Section 2.6.2 Benchmarking and Metrics, we use a case-insensitive WER measure.

Table 14 Example Sentence Benchmark Results

Benchmarked Model	**Example Sentence and Transcription**
Recording	*Ganz korrekt. Wås? Du håst eam doch bestocha, dass er des sågt*[1]
Ground Truth	ganz korrekt was du hast ihn doch bestochen dass er das sagt
End-to-End Results:	
Grosman2021-wav2vec2	ganz korrekt huss du bist jam durch bestorert asdo dist so
Internal-IAIS-wav2vec2	ganz kerringt huss du hast ja doch bestochen dass du das so
Grosman2021-wav2vec2 fine-tuned	ganz korrekt wus du hast ihm doch bestochen dass er das sagt
Internal-IAIS-wav2vec2 fine-tuned	ganz kürekt was du hast ihm doch bestochen dass er das sagt
Hybrid Results:	
Hybrid Baseline	ganz gering Bus gestohlener, dass er das so
AM Training	ganz korrekt Bus du hast dir doch gestochert als er das sagt
LM Training	ganz gering Servus du hast dir doch gestohlen hat also das so
AM+LM Training	Ganz gering was du hast dir doch bestochen also das sagt

In the following, we will distinguish between purely orthographic errors and phonemic[2] errors. We assume they are two different classes of errors that can provide meaningful insight regarding the recognition accuracy. Purely orthographic errors occur when a model chooses a spelling that corresponds to the right phoneme sequence but does not follow the prescribed orthography because the word was not encountered frequently enough (e.g., <Behre> for German *Beere*). On the other hand, we will consider an error phonemic if reading the output of the model does not result in the right phoneme sequence (e.g., <kürekt> /ky'rɛkt/ vs. <korrekt> /ko'rɛkt/). Phonemic accuracy likely does not require specific words to occur in the training data but most likely benefits from minimal overfitting, since orthographic rules can disregard phonemic spelling rules (e.g., *korrekt* could correspond to /'kɔrəkt/). Instead, we assume that a phonemic error can indicate imperfect acoustic modeling. The models that were acoustically trained at some point using Ger-TV1000h tend to fail to recognize the word *korrekt* in this case. Training of the acoustic modeling improved this in the case of the hybrid model and arguably the end-to-end models. The fine-tuned internal-IAIS-wav2vec2 still did not correctly recognize one phoneme and had another spelling error. This shows the capability

[1] Our quasi-traditional Bavarian transcription with German punctuation for better intelligibility.

[2] Here, we follow the linguistic definition of *phoneme*. Our previous argument regarding the usage of *phone* vs. *phoneme* applied to the phonetic representations inside the hybrid model's lexicon and therefore does not apply here.

of end-to-end models to spell unknown words, which may be advantageous or disadvantageous. Another challenging word is (ger.) *was* 'what', pronounced with a back rounded vowel, the phoneme /ɔ ~ o/, whose actual realization here is closer to a stressed close short [o] or even a lax [ʊ], which most models seem to recognize as [ʊ] (*Bus, huss, wus, Servus*). Only the fine-tuned internal-IAIS-wav2vec2 and the hybrid model with combined AM- and LM-training could recognize the word correctly. All models failed to recognize the word form (ger.) *ihn* 'him (accusative)'. However, the fine-tuned internal-IAIS-wav2vec2 and fine-tuned grosman2021-wav2vec2 could recognize the right word — more precisely *lexeme* (Lyons, 1963, p. 12) — and only confused two inflection forms. This can be seen as an appropriate acoustic recognition. Any better recognition via acoustic modeling could be considered overfitted since we consider the final *-m* to be phonetically unpredictable. A dedicated LM would make such an error unlikely.

5 Conclusion

In this work, we explored different approaches to improving the ASR for the German dialect Bavarian for which little transcribed training data exists. The aim was to transcribe Bavarian dialect to written Standard High German.

From the annotated data we could gather, we created datasets for training, validation, and testing. This resulted in approximately 24 hours of annotated training data, while the three test sets amount to approximately 6.7 hours in total.

Our Bavarian models were based on the Fraunhofer IAIS hybrid model and two end-to-end models based on wav2vec 2.0. The experiments in this work may be classified using two categories: Rule-based approaches and data-driven approaches. While the wav2vec 2.0 models were trained in a purely data-driven manner on approximately 24 hours of Bavarian data, both approaches were applied to the hybrid model. The data-driven approach for the hybrid model also included AM training on approximately 24 hours of Bavarian speech. Additionally, an LM was trained on a larger set of transcriptions and subtitle data: a text corpus of annotations from the annotated training set and subtitles of a Bavarian television series. The rule-based approach was conducted via lexicon modifications using phonetic mapping rules. These rules were based on linguistic knowledge.

In our work, we decided on three research questions (see 1.4 Objective and Research Questions) which will be discussed in the following:

Q1: *How can Bavarian ASR be improved on Standard German ASR systems?*

To explore Q1, we conducted experiments on rule-based phonetic lexicon additions and data-driven training. We found that rule-based and data-driven approaches are suitable for improving Bavarian ASR. The improvements gained by data-driven training are, however,

much more substantial. The end-to-end training in this work was especially successful, although language modeling affects its robustness. All in all, the data-driven approaches showed similar trends of improvement and challenges. Nevertheless, they are not comparable to the rule-based approach since we were forced to leave the creation of a combined model with all successful rules to future work.

We found that phonetic mapping rules for vowels tend to improve the WER slightly, while consonantal mapping rules are more challenging to implement or not beneficial. Combining all successful rules might increase the effect of all included rules since the correct recognition of many words is limited by more than one phone. While the creation of the rules and their combination to one unified model were challenging factors, we were able to create a workflow, formalize our procedure, and identify challenges that must be considered when creating one unified model. This enables the efficient creation of new rules and their combination in the future.

Regarding data-driven training, we found that the baseline models previously trained with Ger-TV1000h were substantially better-suited for Bavarian recognition than grosman2021-wav2vec2 which was trained on the Common Voice dataset. Using approximately 24 hours of mostly Bavarian dialect broadcast data greatly improved the acoustic modeling of the models. On the other hand, training may affect the language modeling performance in different ways for the different test sets. This seems to affect end-to-end models across different test sets similarly to the language model training. End-to-end systems, while they could reach the best performance in many cases, may be less robust when using the procedure in this work due to the lack of an AM-LM split. This might be remediated to some degree by combining the end-to-end model with a dedicated LM.

Both hybrid models as well as wav2vec 2.0 models achieve substantial improvements compared to the Fraunhofer hybrid German model. Wav2vec 2.0 produces the best model for two out of three test sets, however, it has disadvantages regarding language modeling. The LM training procedure in this work did not produce a robust model due to its small training size but showed unexpectedly high improvements to both the perplexity and WER on one test set.

Our findings suggest that data-driven training of the acoustic modeling via transfer learning shows the greatest potential for improving Bavarian speech recognition despite the limited training data. For higher improvements, it can be combined with language model training.

Q2: *How can the limited amount of Bavarian data be utilized effectively for ASR model creation and evaluation?*

Due to the limited test data, we decided to split our largest dataset — which was originally intended for training — into a training set, a validation set, and a test set. Since we achieved significant improvements in the acoustic modeling with the smaller training set, we consider the split a success. Other data-sources were also gathered, but could not replace this split. By using the split of the largest set and the other data-sources, three test sets were created. These were kept separate to offer more information and reveal biases and overfitting. The different test sets helped to show the different effects of acoustic modeling and language modeling before and after training. The improvements to the acoustical modeling were high and applied to all test sets. This shows that the training data was sufficient for meaningful training and that the learned knowledge was not overfitted to the largest test set.
Furthermore, since LM training does not require aligned audio data, the training data could be supplemented using subtitles from a Bavarian television series to create a dataset for LM training. LM training was unexpectedly beneficial on one test set, but not robust. This suggests benefits when using such data for interpolated LMs.

Q3: *Can a model be created that can reduce the WER below the 40% mark?*

While one of the test sets tended to show deteriorations in some conditions, we could reach lower than 40% WER on the largest test set using data-driven approaches for our wav2vec 2.0 and hybrid models.

The findings differ across the test sets due to substantial differences in the data. Particularly language modeling is affected differently by our experiments. This could be seen in the smallest dataset “Bairisch-2020” achieving WERs lower than 40% for models only trained with Ger-TV1000h (hybrid WER 39.59%, end-to-end WER 36.00%). Acoustic model training further reduced the hybrid model WER to 33.19%. This small dataset was negatively affected by language model training.

WERs below 40% could be achieved on the largest test set “Oachkatzl-test” with the data-driven approaches in this work. Training both the LM and the AM of the hybrid model led to a WER of 39.33%. The fine-tuning of Internal-IAIS-wav2vec2 — which due to its end-to-end nature also involved fine-tuning of the language modeling — brought the WER down to 36.91%. Therefore, we can infer that, given our data, training of both the acoustic modeling and language modeling should be considered to achieve WERs as low as 40%.

6 Outlook

Since the aim of this work was exploratory, we will conclude this work with an outlook of possible future steps and promising approaches given our findings.

The rule-based approach in this work requires a combination of all successful rules to one unified model. Furthermore, additional phones can be tested using the developed workflow. Apart from the application of the rules in this work, we view the developed process as a suitable means of generating G2P training data for dialectal spellings that regular German G2P models fail to correctly transcribe. Using dialectal spellings and the trained G2P model, dialectal lexicon additions might be based on spellings.

With regard to the data-driven approaches, attempts to increase the usable training data should be considered. The results show that even a small number of hours may substantially improve the performance. For this, Austrian sources may be evaluated. Our findings indicate that a language model should be created based on more “Dahoam is Dahoam” subtitle data in conjunction with Standard German or subtitles from regional Bavarian broadcast. We assume this will increase the robustness of the model while improving the performance on a subset of Bavarian speech. The acoustic model training could be combined with rule-based lexicon additions before training. Also, as Gref personally suggested after the initial version of this work, it may be beneficial to experiment with leaving out the i-vectors that enable speaker recognition from the input features. Due to the little choice in Bavarian speech data, there are segments that have multiple speakers. We did not anticipate this in our experiments, but the speaker recognition might also affect the ASR performance when multiple speakers occur in one segment.

Finally, the end-to-end approach delivered excellent results and our findings suggest striking possibilities for further improvement. We found that our end-to-end approach using wav2vec 2.0 models is likely to benefit from lowering the learning rate. Apart from this, we plan to combine it with a language model. While this language model will have a smaller effect than for the hybrid model, we expect it to drastically reduce some minor spelling errors, inflection errors and some cases of wrong word choice that affect the WER even though the errors are subtle. Also, an LM can enable the model to distinguish casing. The combination of a Bavarian wav2vec 2.0 model with an LM will profit from our findings regarding LM training. Lastly, one option remains that might prove more advantageous than any other: A wav2vec 2.0 model can be trained for a combined task of ASR and dialect classification. Using multiple varieties, including forms of Bavarian, Swiss German, and Standard German, can drastically increase the amount of training data drastically, while the classification task will contribute to an understanding of variety-specific phone correspondences and grammatical differences.

Bibliography

Adler, A., & Plewnia, A. (2020). Aktuelle Bewertungen regionaler Varietäten des Deutschen. Erste Ergebnisse der Deutschlanderhebung 2017. *Regiolekte: Objektive Sprachdaten und subjective Sprachwahrnehmung*, 15-36. Tübingen, Germany: Narr Francke Attempto. doi:10.2357/9783823393177-01

Ali, A., Nakov, P., Bell, P., & Renals, S. (2017, Dec). WERD: Using social text spelling variants for evaluating dialectal speech recognition. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 141–148. doi:10.1109/ASRU.2017.8268928

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., . . . Weber, G. (2020, May). Common Voice: A Massively-Multilingual Speech Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4218-4222. Marseille, France: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/2020.lrec-1.520

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 33*, 11259-11272. (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Eds.) Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/81e3225c6ad49623167a4309eb4b2e75-Paper.pdf

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983, May). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and*, 179-190.

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication, 50, 5*, 434–451. doi:https://doi.org/10.1016/j.specom.2008.01.002.

Brockhaus, W. (2012). Final Devoicing in the Phonology of German. (G. Müller, & K. Heusinger, Eds.) Tübingen, Germany: Max Niemeyer Verlag. Retrieved from https://www.degruyter.com/document/doi/10.1515/9783110966060/html

Cattette. (2020, May 22). A map of the extent of the Bavarian language. *Austro_Bavarian_Languages-01.png*. Retrieved 11 15, 2022, from https://en.wikipedia.org/wiki/File:Austro_Bavarian_Languages-01.png

Chomsky, N., & Halle, M. (1968). The Sound Patterns of English. New York, Evanston, and London: Harper & Row.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *Proc. Interspeech 2021*, 2426–2430. doi:10.21437/Interspeech.2021-329

ELAN (Version 6.3) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://archive.mpi.nl/tla/elan

Federico, M., Bertoldi, N., & Cettolo, M. (2008, Sept). IRSTLM: An open source toolkit for handling large scale language models. 1618–1621. doi:10.21437/Interspeech.2008-271

Furhop, N., & Peters, J. (2013). Einführung in die Phonologie und Graphematik. Stuttgart, Weimar, Germany: Metzler.

Garner, P. N., Imseng, D., & Meyer, T. (2014). Automatic speech recognition and translation of a Swiss German Dialect: Walliserdeutsch. *Interspeech 2014*. doi:DOI:10.21437/Interspeech.2014-480

Gref, M. (2022, Oct). Robust Speech Recognition via Adaptation for German Oral History Interviews. Rheinische Friedrich-Wilhelms-Universität Bonn. Retrieved from https://hdl.handle.net/20.500.11811/10373

Gref, M., Schmidt, C., Behnke, S., & Köhler, J. (2019). Two-staged acoustic modeling adaption for robust speech recognition by the example of german oral history interviews. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 796–801. doi:https://doi.org/10.1109/ICME.2019.00142

International Phonetic Association. (2015). Handbook of the International Phonetic Association. Cambridge University Press.

IPA Chart. (2020). IPA Chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2020 International Phonetic Association. Retrieved from https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/IPA_charts_E.html

König, W. (2011). dtv-Atlas - Deutsche Sprache. Deutscher Taschenbuch Verlag.

Levenshtein, V. I. (1966, Feb). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10 (8)*, 707-710.

Lyons, J. (1963). Structural semantics: an analysis of part of the vocabulary of Plato.

Max Planck Institute for Evolutionary Anthropology. (2014). The Generic Style Rules for Linguistics. Retrieved Nov 15, 2022, from https://www.eva.mpg.de/linguistics/past-research-resources/resources/generic-style-rules/

Moore, R. K., & Skidmore, L. (2019). On the Use/Misuse of the Term 'Phoneme'. *Interspeech 2019*, 2340–2344. doi:http://dx.doi.org/10.21437/Interspeech.2019-2711

Ney, H. (1990). Acoustic modeling of phoneme units for continuous speech recognition. *Signal Processing V: Theories and Applications, Fifth European Signal Processing*

Conference, 65–72. (E. M. L. Torres, Ed.) Barcelona, Spain: Elsevier Science Publishers B. V.

Ney, H., & Ortmanns, S. (1999). Dynamic programming search for continuous speech. *IEEE Signal Processing Magazine, 16(5)*, 64–83. Retrieved from https://ieeexplore.ieee.org/document/790984

Niebaum, H., & Macha, J. (2014, Aug). Einführung in die Dialektologie des Deutschen. De Gruyter. doi:https://doi.org/10.1515/9783110338713

Nigmatulina, I., Kew, T., & Samardzic, T. (2020, Dec). ASR for non-standardised languages with dialectal variation: the case of Swiss German. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 15–24. Barcelona, Spain: International Committee on Computational Linguistics (ICCL). Retrieved from https://aclanthology.org/2020.vardial-1.2

Peddinti, V., Wang, Y., Povey, D., & Khudanpur, S. (2018). Low latency acoustic modeling using temporal convolution and LSTMs. *IEEE Signal Processing Letters, 25(3)*, 373-377. doi:10.1109/LSP.2017.2723507

Pike, K. L. (1943). A critical analysis of phonetic theory and a technic for the practical description of sounds (= University of Michigan publications. Language and literature. *Volume 21*. (A. Arbor, Ed.) University of Michigan Press.

Pittner, K. (2013). Einführung in die germanistische Linguistik. Darmstadt, Germany: WBG.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The kaldi speech recognition toolkit. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., . . . Khudanpur, S. (2016). Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. *Proc. Interspeech 2016*, 2751-2755. doi:10.21437/Interspeech.2016-595

Pritzen, J. (2021). Improving Grapheme-to-Phoneme Conversion for Anglicisms in German Speech Recognition. Retrieved from https://publica.fraunhofer.de/entities/publication/c351f92c-a0b2-4ad9-bc6f-cec9d1c05d47/details

Schmidt, C., Stadtschnitzer, M., & Koehler, J. (2016). The Fraunhofer IAIS Audio Mining System: Current State and Future Directions. *Speech Communication; 12. ITG Symposium*, 1–5. Retrieved from https://ieeexplore.ieee.org/document/7776158

Stadtschnitzer, M. (2018, Oct). Robust Speech Recognition for German and Dialectal Broadcast Programmes. Retrieved from https://bonndoc.ulb.uni-bonn.de/xmlui/bitstream/handle/20.500.11811/7658/5236.pdf

Stadtschnitzer, M., & Schmidt, C. (2015). Implementation of a live dialectal media subtitling system. *Proc. Interspeech 2015*, 728–729. Retrieved from https://www.isca-speech.org/archive_v0/ interspeech_2015/i15_0728.html

Stadtschnitzer, M., & Schmidt, C. (2018, May). Data-driven pronunciation modeling of Swiss German dialectal speech for automatic speech recognition. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1498

Stadtschnitzer, M., Schwenninge, J., Stein, D., & Köhler, J. (2014, April). Exploiting the large-scale German broadcast corpus to boost the fraunhofer IAIS speech recognition system. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3887–3890. Reykjavík, Iceland: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/858_Paper.pdf

Sushmita, C., Vijayshri, C. N., & Challa, K. A. (2016). Statistical and linguistic knowledge based speech recognition system: Language acquisition device for machines. *Information Systems Design and Intelligent Applications*, 613–619. (S. C. Satapathy, J. K. Mandal, S. K. Udgata, & V. Bhateja, Eds.) New Delhi, India: Springer India. Retrieved from https://link.springer.com/ chapter/10.1007/978-81-322-2752-6_60

Trubetzkoy, N. S. (1939). Grundzüge der Phonologie. Prague, Czech Republic. Retrieved from https://pure.mpg.de/rest/items/item_2399346/component/file_2399345/content

Uppstad, P. H., & Tønnessen, F. E. (2010). The Status of the Concept of 'Phoneme' in Psycholinguistics. *Journal of Psycholinguistic Research 39*, 429–442. Retrieved from https://link.springer.com/article/10.1007/s10936-010-9149-8

Wang, D., & Zheng, T. F. (2015). Transfer learning for speech and language processing. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1225–1237. doi:https://doi.org/10.1109/APSIPA.2015.7415532

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems, Part IV, section B*. (D. Gibbon, R. Moore, & R. Winski, Eds.) Berlin and New York: Mouton de Gruyter. doi:10.1007/s10936-010-9149-8

Wester, M., & Kessens, J. M. (2000, Sept). Using Dutch Phonological Rules to Model Pronunciation Variation in ASR. Retrieved from https://www.researchgate.net/publication/2470091_Using_Dutch_Phonological_Rules_to_Model_Pronunciation_Variation_in_ASR

Wiesinger, P. (1980). ”Sprache”, ”Dialekt” und ”Mundart” als sachliches und terminologisches Problem. *Ergebnisse d. Internationalen Symposions "Zur Theorie d. Dialekts," Marburg/Lahn, 5.-10. September 1977*, 177–194. (Göschel/Ivić/Kehr, Ed.)

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2020, oct). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-demos.6

Zehetner, L. (1985). Das bairische Dialektbuch. München, Germany: Beck. Retrieved from https://epub.uni-regensburg.de/25998/

List of Software and Toolkits

ELAN — Used in the initial phases for viewing transcribed data. (ELAN (Version 6.3) [Computer software], 2022). https://archive.mpi.nl/tla/elan

IRSTLM — Used to create n-gram models. https://ict.fbk.eu/irstlm/

Kaldi — Used to create ASR models. http://kaldi-asr.org/

Sequitur G2P — The baseline lexicon was originally created using Sequitur G2P. https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html