Julian Linke

What's so complex about conversational speech?

Prosodic prominence and speech recognition challenges

Dipl.-Ing. Julian Linke, BSc

What's so complex about conversational speech? Prosodic prominence and speech recognition challenges

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisors

Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin

Ass. Prof. Mag.rer.nat. dr. Barbara Schuppler

External Reviewer

Dr. L.F.M. Louis ten Bosch

Signal Processing and Speech Communication Laboratory

Graz, 2025

Abstract

This thesis presents the analysis and evaluation of acoustic representations and models for conversational speech for two tasks: prosodic prominence classification and automatic speech recognition (ASR). Conversational speech poses unique challenges compared to read or prepared speech due to characteristics such as lively turn-taking, incomplete utterances, disfluencies, and high degree of pronunciation variation. Given these characteristics, both prosodic annotation tools and ASR systems trained on the typical benchmark datasets perform significantly worse on conversational speech. This thesis thus follows two aims, 1) to analyze acoustic representations for conversational speech using explainable machine learning (ML) methods, and 2) to improve the performance of prosodic prominence classification and ASR systems, as measured with standard performance measures. Our experiments on prosodic prominence classification revealed that the main acoustic cues for perceived prominence were the durational features. We introduce novel entropy-based prosodic features, which showed to encode necessary durational information along with information on pitch and loudness, leading to detection performances which aligned with inter-annotator agreements for the different prominence levels. These entropy-based prosodic representations were further used to examine their impact on utterance-level word error rates (WERs) of HMM- and transformer-based ASR systems. Our results reveal significant effects of durational and prosodic features on WER, but also how they interact with pronunciation variation and utterance-level complexity measures. Finally, we developed prominence detectors and prominence-aware ASR systems and explored how prosodic information is encoded through fine-tuning of self-supervised speech representations, indicating the feasibility of integrating prosodic information into ASR. Given that our experiments were based on data from conversational Austrian German, we had to deal with high variation stemming from dealing with a (low-resourced) regional variety of a (well-resourced) language in addition to the high variation between speakers and between different speaker pairs given the casual speaking style. Using clustering methods for shared discrete speech representations we demonstrated their effectiveness in differentiating language and variety aspects and capturing speaker differences across styles. The distances between quantized latent speech representations showed to meaningfully capture fine-grained differences between speakers when producing different speaking styles. Overall, this thesis provides insights into the complexities of conversational speech and demonstrates how the analysis and evaluation of acoustic representations and models deepen our understanding of conversational speech. The findings have implications for various applications such as human-machine interaction, conversation transcription and hearing aid technology.

Kurzfassung

Diese Dissertation präsentiert die Analyse und Evaluierung akustischer Repräsentationen und Modelle für Konversationssprache in zwei Aufgabenbereichen: der Klassifikation prosodischer Prominenz und der automatischen Spracherkennung (ASR). Konversationssprache bietet im Vergleich zu gelesener oder vorbereiteter Sprache einzigartige Herausforderungen und ist durch eine lebhafte Gesprächsdynamik, unvollständige Äußerungen, Füllwörter und ein hohes Maß an Aussprachevariation charakterisiert. Aufgrund dieser Merkmale funktionieren sowohl prosodische Annotations-Tools als auch ASR-Systeme, die auf typischen Benchmark-Datensätzen trainiert wurden, bei Konversationssprache signifikant schlechter. Diese Dissertation verfolgt zwei Ziele: 1) die Analyse akustischer Repräsentationen für unter Verwendung erklärbarer Methoden des maschinellen Lernens (ML), und 2) die Verbesserung der Leistung von Systemen zur Klassifikation prosodischer Prominenz und ASR-Systemen, gemessen an standardisierten Evaluierungsparametern. Die Experimente zur Klassifikation prosodischer Prominenz zeigten, dass die wichtigsten akustischen Merkmale für wahrgenommene Prominenz die dauer-bezogenen Merkmale waren. Es wurden entropiebasierte prosodische Merkmale eingeführt, die notwendige Informationen zur Dauer sowie Informationen über Tonhöhe und Lautstärke kodieren. Die dadurch erreichte Modellgenauigkeit stimmte mit jener der menschlichen Annotatoren überein. Diese entropiebasierten prosodischen Repräsentationen wurden weiter verwendet, um ihren Einfluss auf Wortfehlerraten (WERs) von HMM- und transformerbasierten ASR-Systemen zu untersuchen. Die Ergebnisse zeigen signifikante Effekte von Dauer- und Prosodiemerkmalen auf die WER, aber auch, wie sie mit Aussprachevariationen und Komplexitätsmaßen auf Äußerungsebene interagieren. Schließlich präsentiert diese Dissertation Prominenzdetektoren und prominenz-sensitive ASR-Systeme, die neben der orthographischen Transkription auch Prominenzniveaus automatisch annotieren, und untersucht, wie prosodische Information durch das Fine-Tuning von self-supervised Sprachrepräsentationen enkodiert wird. Da alle Experimente dieser Dissertation auf Daten des Deutsch basieren, war es notwendig, Methoden zu entwickeln, die mit hoher regionaler Variation umgehen können - zusätzlich zur hohen Variabilität zwischen Sprecher*innen und Sprecherpaaren aufgrund des informellen Sprechstils. Durch den Einsatz von Clustering-Methoden für self-supervised Sprachrepräsentationen wurde gezeigt, dass diese effektiv Sprache unterschiedlicher Stile und Varietäten unterscheiden. Die Abstände zwischen quantisierten latenten Sprachrepräsentationen erwiesen sich als aussagekräftig bei der Erfassung feingranularer Unterschiede zwischen Sprecher*innen bei der Produktion unterschiedlicher Sprechstile. Insgesamt bietet diese Dissertation Einblicke in die Komplexität der Konversationssprache

und zeigt, wie die Analyse und Bewertung akustischer Repräsentationen und Modelle unser Verständnis von Konversationssprache vertiefen. Die Ergebnisse dieser Arbeit können in verschiedenen Bereichen zur Anwendung kommen, im speziellen um Mensch-Maschine-Interaktion natürlicher zu gestalten, und linguistische Sprachkorpora automatisch auf Wort und Prominenzniveau zu annotieren.

Contents

Acknowledgements

I extend my deepest gratitude to my supervisors who guided me through this journey. To Gernot Kubin for his illuminating scientific discussions, innovative ideas, and remarkable ability to explain complex concepts with clarity, as well as his invaluable support in administrative matters. My heartfelt thanks to Barbara Schuppler, who equipped me with essential scientific tools, from research methodologies to writing skills. Her passion and enthusiasm for research have been truly inspiring. I would like to thank Louis ten Bosch for taking the time to review my thesis and providing valuable feedback that helped enhance the final manuscript.

I'm grateful to my SPSC colleagues - Saskia, Anneliese, and Philipp for engaging discussions, Martin Hagmüller for insightful contributions and Markus for exceptional system administration. I extend my appreciation to all other colleagues at SPSC for enriching interactions.

I'm particularly grateful to Phil Garner's team at Idiap, especially Phil and Abbas, who helped shape my focus on novel speech recognition methods during my research stay in Martigny. I'm thankful for meeting Peter Mihajlik at LREC in Marseille, leading to productive collaborations with him, Katalin, and Mate. Special thanks to Bernhard Geiger from Know Center, whose calm demeanor, deep knowledge, and clear explanations were crucial during the final phase of my PhD.

To my family - my parents Angelika and Steffen for making my studies possible, my siblings Ferdinand, David, and Smilla for their support, and grandmother Marga for her comforting phone conversations. Words cannot express my gratitude to Frauke - this thesis wouldn't exist without you. To Luna, my faithful companion, who spent countless hours by my side during this journey. I'm grateful to my Bamberg friends - Jonas, Christian, and to Markus, who also designed this thesis cover - and my Graz circle - Nicolai, Niklas, Ingo, Zulaa, Robert, and Simon - for all the times we shared. In loving memory of my grandparents Erwin, Karola, and Lothar, and my dear friend Juan - you remain part of my best moments.

Nomenclature

#phones	Number of realized phones per utterance
#tokens	Number of word tokens per utterance
PL0	Words annotated with no prominence level
PL12	Words annotated with weak, strong and empathic prominence level
PL1	Words annotated with weak prominence level
PL2	Words annotated with strong and empathic prominence level
Accuracy	Proportion of correctly predicted observations to the total number observations [%]
AR	Articulation rate of an utterance [s^{-1}]
ASR	Automatic speech recognition
BEA	BEA database
BECS	BEA spontaneous speech component
BERS	BEA read speech component
Chap.	Chapter
CS	Conversational speech
CV	Cross-validation
DT	Decision tree classifier
DUR	Durational features
ENT	Entropy-based features
Eq.	Equation

F0	Fundamental frequency (features); refers to either F0 contours or the broader category of F0-related features, depending on context
F1-score	Balanced F-score: Harmonic mean of Precision and Recall [%]
Fig.	Figure
GECO	GECO corpus
GEMO	GECO-Mono (unimodal setting of GECO corpus)
GEMU	GECO-Multi (multimodal setting of GECO corpus)
GRASS	GRASS corpus
GRCS	GRASS conversational speech component
GRRS	GRASS read speech component
GSG	German Standard German
HF0N	Normalized conventional entropy of F0 contour
HPSF0	Pseudo-entropy of F0 contour
HPSF0N	Normalized pseudo-entropy of F0 contour
HPSRMS	Pseudo-entropy of RMS contour
HPSRMSN	Normalized pseudo-entropy of RMS contour
HRMSN	Normalized conventional entropy of RMS contour
Kaldi	ASR architecture/system Kaldi; trained exclusively on low-resourced conversational Austrian German speech data (unless specified otherwise) with either cross-entropy or LF-MMI criterion (depending on context)
KICS	KIEL spontaneous speech components (Verbmobil and Videtask)
KIEL	KIEL corpus
KIRS	KIEL read speech component
KIVM	KIEL spontaneous speech component (Verbmobil)
KIVT	KIEL spontaneous speech component (Videotask)
LR	Low-resourced
ML	Machine learning

PHN-LR	Pre-trained and fine-tuned wav2vec2.0 model exclusively on low-resourced Austrian German conversational speech data with character-based target set
PHN-LR	Pre-trained and fine-tuned wav2vec2.0 model exclusively on low-resourced Austrian German conversational speech data with phone-based target set
PHN-XLSR	Fine-tuned wav2vec2.0 cross-lingual speech representation model with character-based target set
PHN-XLSR	Fine-tuned wav2vec2.0 cross-lingual speech representation model with phone-based target set
pplAGS	Language model perplexity of an utterance with respect to a trigram trained with approx. 220k Austrian German sentences
pplWIKI	Language model perplexity of an utterance with respect to a four-gram trained on 5M German sentences
Precision	Positive predictive value: Fraction of relevant observations among all retrieved observations [%]
PronD	Pronunciation difference reflecting the degree of reduction of an utterance
PronLD	Pronunciation Levenshtein distance reflecting the degree of reduction and deviation from the standard pronunciation of an utterance
Recall	Sensitivity: Fraction of relevant observations among all relevant observations [%]
RFC	Random forest classifier
RMS	Root mean square (features); refers to either RMS contours or the broader category of RMS-related features (depending on context)
RS	Read speech
Sec.	Section
Tab.	Table
UttDur	Total duration of an utterance [s]
w2v	Fine-tuned wav2vec2.0 ASR system with greedy decoder; no lexicon/LM
w2vLM	Fine-tuned wav2vec2.0 ASR system with beam-search decoder; includes lexicon/LM

wav2vec2	ASR architecture wav2vec2.0
WER	Word error rate [%]
Whisper	ASR architecture/system Whisper; large-v2 model in zero-shot mode
XLSR	Cross-lingual speech representation model based on the wav2vec2.0 architecture (pre-training with 56000h of multilingual speech data)

Chapter 1

Introduction

This thesis explores conversational speech by analyzing and evaluating acoustic representations and models for prosodic prominence detection and speech recognition. In general, conversational speech refers to a speaking style where two or more people are conducting a spontaneous conversation, resulting in a less structured form than read or prepared speech. Anything beyond these general properties could refer to the *casualness* of the conversational setting, the *relationship* between the conversational interlocutors or the setting of a *face-to-face* conversation. With respect to its structural properties, the conversational speaking style is characterized by a lively turn-taking, resulting in short utterances, grammatically incomplete utterances, self-interruptions, backchannels and disfluencies. Another characteristic is a high degree of pronunciation variation, resulting from acoustic reduction processes and (dialectal) phonological processes (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014). This list of conversational characteristics is by no means complete, but clearly shows that these aspects introduce elusive complexities which need to be taken into account for the development of appropriate applications. This thesis investigates certain aspects of these complexities in conversational speech, which can be beneficial for improving various applications such as human-machine interaction systems (e.g., social robots or speech agents), transcription software for spontaneous conversations between two or more humans (e.g., meeting recordings) or hearing aid technology.

The technology of all these applications in the focus of this thesis is automatic speech recognition (ASR). ASR performance varies greatly across languages (i.e., well-resourced vs. low-resourced) and speaking styles (i.e., read or well-prepareed vs. spontaneous speech) making it challenging to establish a universal benchmark. As an example, there exists the common English read speech corpus LibriSpeech (Panayotov et al., 2015) for which impressive word error rates (WERs) of 1.4 % were reported (Zhang et al., 2020). In contrast, Xu et al. (2021) reach a WER of 12.3 % for read speech corpora in German. Speech recognition results for spontaneous or conversational speech are even more diverse. For instance, ASR results on the Switchboard corpus (J. J. Godfrey et al., 1992), a corpus of spontaneous telephone conversations in American English, demonstrate the progress made in speech recognition with recently reported WERs in the range of 4.3 % (Tüske et al., 2021) to 5.1 % (Xiong et al., 2018), whereas on the same corpus WERs were in the range of 11.5 % to 14.5 % only 10 years ago. These high performances on the

well-resourced language American English, however, can by far not be achieved for low-resource languages, low-resource language varieties nor for more spontaneous speaking styles. For instance, the OpenASR21 challenge showed that for low-resource languages WERs fall in the range of 32 % (Swahili) to 68 % (Farsi) (Peterson et al., 2022). Overall, it is not straightforward to define sub-categories of spontaneous speaking styles which are subject to many factors (cf. Fig. 2.1). Therefore, this makes it difficult to directly compare WERs of different spontaneous speech corpora. One focus of this thesis is the analysis of the conditions under which different ASR architectures (HMM-based vs. transformer-based) face challenges. Additionally, it presents how a combination of data-choice and linguistic knowledge integration improves ASR performance and it shows how self-supervised speech representations are related to speaking styles and language varieties.

Another important aspect of conversational speech is prosodic variation, where in this thesis the focus is specifically on prosodic prominence. Prosodic prominence is a complex phenomenon (B. Wagner et al., 2015) and generally defined as a linguistic entity which stands out from its environment due to prosodic characteristics (Terken & Hermes, 2000). This definition emphasizes the relative nature of prosodic prominence which also includes monosyllabic utterances because they stand out from silence (Terken & Hermes, 2000). Either way, acoustic cues to perceived prominence have not only been analyzed on syllable-level (Kochanski et al., 2005; Mixdorff et al., 2015; Terken & Hermes, 2000; P. Wagner, 2005) but also on word-level (Bishop, 2012; Cole et al., 2019, 2010; Turnbull et al., 2017) or vowel-level (Baumann et al., 2016). The annotation of prominence can be based on either two or more levels (like in this thesis) or on a continuous scale. Surprisingly, the findings in the literature do not suggest a consistent view on which acoustic cues that mainly contribute to the perception of prosodic prominence. This inconsistency may stem from the fact that the perception of prominence is influenced by various factors that go beyond prosodic acoustic cues such as the lexical context or the speaking style. This thesis analyzes which acoustic features contribute how strongly to the perception of prosodic prominence in conversational Austrian German and shows methods to enhance its classification and detection performance.

Applications including ASR and prosodic prominence have strongly been influenced by the recent advancements in the field of machine learning (ML). Initially, ML problems were solved with relatively simple models (e.g., simple linear models or decision trees) whose decision processes were easily interpretable. They were, however, largely replaced by less interpretable deep learning models (Goodfellow et al., 2016). At present, especially due to the advances in computing power, it is possible to train highly parameterized models (e.g., ChatGPT has 175 billion parameters and its successors could be six times larger (M. Mijwil et al., 2023)). While these models can solve increasingly challenging problems, their complexity comes at the cost of interpretability for scientists employing them. Interpreting these more complex models remains feasible especially when they incorporate model-specific analysis methods (e.g., random forests with impurity-based feature importances (Breiman, 2001, 2002); cf. Chap. 3). Simultaneously, there exists a multitude of interpretable model-agnostic methods (e.g., SHAP (Lundberg & Lee, 2017) or permutation feature importances (Breiman, 2001, 2002; Fisher et al., 2019)) which aim to provide interpretations for various model types. These methods are not only theoretically established, but are also becoming increasingly accessible due to rapidly evolving

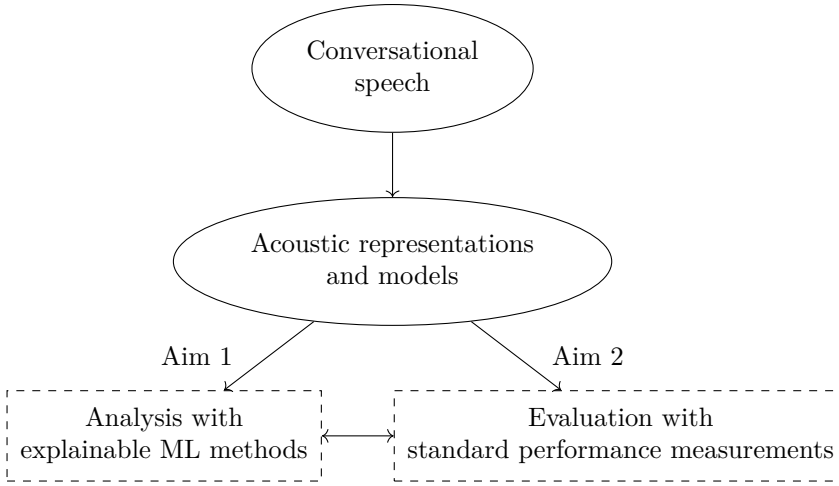


Figure 1.1: Overview of the two research aims (ellipses describe entities and dashed rectangles processes): **Conversational speech** builds the origin for most experiments described in this thesis and it generally includes acoustic data together with meta data. The next step involves the processing of the speech data in **acoustic representations and models** (i.e., training or fine-tuning of word or word sequence models as well as encoding with speech representation models). This leads to the aims of this thesis which comprise 1) the **analysis of acoustic representations and models with explainable machine learning methods** (e.g., tree-based or feature transformation methods) and 2) the **evaluation of acoustic representations and models with standard performance measurements** (e.g., accuracies/recalls/F1-scores or WERs). These research aims lead to specific research questions which are discussed in this thesis in order to deepen our understanding of conversational speech characteristics. The double-headed arrow between the dashed rectangles indicates that the two research aims depend on each other.

research in the field of explainable AI. Hence, there is always a necessity to search and apply new interpretable model-agnostic or model-specific perspectives. In this thesis, modern tree-based models with advanced model-specific methods are utilized, enabling the analysis of interacting variables (e.g., Interaction Forests (Hornung & Boulesteix, 2022a); cf. Chap. 4.2). Furthermore, more advanced deep neural network architectures which learn powerful latent or contextualized representations (i.e., wav2vec2.0 (Baevski, Zhou, et al., 2020)) are analyzed with respect to feature transformations and clustering methods in order to provide a better interpretability of the information encoded in these representations (cf. Sec. 4.4).

1.1 Research aims

This thesis has two main research aims (cf. Fig. 1.1): The first aim is the analysis of acoustic representations and models with explainable tree-based or feature transformation methods. The second aim is the evaluation of acoustic representations and models with standard performance measurements. Both research aims are investigated on two different tasks for conversational speech: prosodic prominence classification and automatic speech recognition. All studies presented contribute to deepen our understanding of conversational speech characteristics.

1.1.1 Aim 1: Analysis of acoustic representations for conversational speech with explainable machine learning methods

The first aim of this thesis is the *analysis* of acoustic representations and models with *explainable* tree-based or feature transformation *methods* in order to deepen our understanding of conversational speech characteristics.

To achieve this aim, we investigated the following four research questions:

RQ1: Which are the main acoustic cues for prosodic prominence?

RQ2: Are WERs of conversational speech affected by utterance-level features?

RQ3: What do shared discrete speech representations encode with respect to language varieties, speaking styles and speakers?

RQ4: Does the fine-tuning of self-supervised speech representations implicitly encode prosody?

1.1.2 Aim 2: Evaluation of acoustic representations and models for conversational speech with standard performance measurements

The second aim of thesis is the *evaluation* of acoustic representations and models with standard *performance* measurements in order to deepen our understanding of conversational speech characteristics

Related to this aim, we investigate the following two research questions:

RQ5: Are word-level prominence classification results with prosodic features or word-level prominence detection results with fine-tuned speech representations in line with inner-annotator agreements?

RQ6: How do low-resourced HMM-based ASR systems compare to low-resourced or data-driven transformer-based ASR systems in terms of effectiveness for recognizing Austrian German conversational speech?

1.2 Contributions and outline

The core of this thesis is based on six reformatted articles[1] (incorporated into Chap. 2 - Chap. 4), which encompass the two research aims with the corresponding research questions. The first aim is to analyze acoustic representations and models for conversational speech, such as classifying prosodic prominence or performing automatic speech recognition, by using explainable methods like tree-based algorithms or feature transformation techniques. The second aim is to evaluate these acoustic representations and models using standard performance metrics, including accuracies/recalls/F1-scores or WERs. In general, this thesis has three core chapters: Chap. 2 presents conversational speech resources and initial ASR experiments on these resources, Chap. 3 presents experiments for prosodic prominence classification and Chap. 4 ASR experiments. For each of these chapter, we briefly outline their main contributions here:

Chap. 2 (cf. Linke, Wepner, et al. (2023)) provides a categorization of spontaneous speech corpora with respect to critical questions like *task-oriented?*, *experimenter present?*, *number of speakers?*, *casual?*, *relationship?* or *face-to-face?* (cf. Sec. 2.1) and it also describes in more detail speech corpora with spontaneous/conversational speech components which are relevant for this thesis (cf. Sec. 2.2). This chapter also introduces initial (low-resourced) speech recognition experiments for Austrian German read and conversational speech using Kaldi (cf. Sec. 2.3). These Kaldi-based experiments serve as a baseline, demonstrating the challenges of ASR for conversational Austrian German. Thus, this chapter is meant as an extension of the motivation of this thesis by describing the research gap that requires a more differentiated view of experiments with conversational speech data.

Chap. 3 investigates word-level prominence classification for read and conversational speech. The evaluation is based on different training/test conditions or specific feature selections. Furthermore, prosodic features are analyzed with respect to χ^2-statistics or impurity-based feature importances which are derived from random forest classification models (cf. Sec. 3.2; Linke et al. (2020)). In addition, novel entropy-based prosodic features are introduced to the field of prosody (cf. Sec. 3.3; Linke, Kubin, and Schuppler (2023)). Overall, the contributions of this chapter show that, with minor exceptions, the classification of word-level prominence is consistent with the inter-annotator agreement and that word duration is by far the most important feature.

Chap. 4 includes all ASR experiments with a focus on conversational speech. The first study (cf. Sec. 4.2; Linke et al. (2024)) provides insights into the challenges of speech recognition with conversational speech when comparing different automatic speech architectures (i.e., Kaldi, wav2vec2 and Whisper). More specifically, four different ASR systems are examined in more detail with respect to three aspects: 1) A comparison of HMM-based and transformer-based architectures, 2) the influence of the amount of training data from the target language and style, and 3) the meaning of the incorporation of explicit linguistic knowledge. One major finding of this study

[1]The full list of the publications included in this thesis is given in the Curriculum Vitae (cf. page 141). Throughout the thesis, the publications are of course referenced at the relevant places. In all included publications, my contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

is that with zero-shot learning, the performance on out-of-domain conversational speech is poor especially for *short* utterances and large pronunciation variation. The second study deals with the role of data in conversational speech recognition (cf. Sec. 4.3; Linke et al. (2022)). In particular, I show that a low-resource language processing assumption is permissible for the conversational speech component of the GRASS corpus. In addition, I present also a deeper discussion on the role of linguistic knowledge, the role of targets and the difference between inter-conversation and inter-speaker variation. Both studies (cf. Sec. 4.2 and Sec. 4.3) reveal that transformer-based architectures pre-trained on large amounts of data outperform HMM-based architectures if fine-tuned or trained on the same corpus. Likewise, if the decoder of these fine-tuned transformers includes linguistic knowledge in the form of a lexicon or language model, the performance is generally better and at the same time more robust against acoustic and lexical variation. Next, we present an analysis of self-supervised speech representations with respect to languages, varieties, speaking styles and speakers (cf. Sec. 4.4; Linke, Kadar, et al. (2023)). This study reveals that the calculation of distances with respect to shared quantized latent speech representations is also meaningful on a much finer granularity level (i.e., per speaker per speaking style instead of only per languages). The last experiment presented in this thesis (cf. Sec. 4.5) presents yet unpublished work on integrating prominence detection into a ASR by introducing *prominence-aware ASR*. More precisely, I show that performance of word-level prominence detection (in contrast to prominence classification; cf. Chap. 3) can be integrated into a transformer-based automatic speech recognition framework.

Chap. 5 discusses all findings in the light of the two overarching research aims and presents the contributions corresponding to the detailled research questions (**RQ1** - **RQ6**). This chapter presents ideas for future work related to each aim and concludes this thesis.

Chapter 2

Conversational speech resources

This chapter provides a categorization of spontaneous speech corpora and describes the speech materials relevant for this thesis in detail. Additionally, this chapter presents initial speech recognition experiments for Austrian German with the speech recognition toolbox Kaldi (Povey et al., 2011) to obtain a first impression of the performance of ASR on read vs. conversational speech. Experiments with more recent ASR systems are presented later in Chap. 4.

2.1 Categorization of spontaneous speech corpora

In ASR literature, speaking styles are defined after different criteria, and terms for corpus categorization such as "read", "spontaneous" and "conversational" may actually point to corpora of very different characteristics. Fig. 2.1 shows a categorization scheme that helps us describing the style of the corpora we use in this study, and in general, helps us defining which of the corpora widely used in the ASR community are actually comparable to each other. Note that we do not present a full categorization of all possibilities in Fig. 2.1, but end the tree at those points, where the corpora used in this study drop out.

In general, we can distinguish read from spontaneous speech, where spontaneous contrasts from read given that lexemes and their word order are planned spontaneously. Examples for read speech (RS) are LibriSpeech (Panayotov et al., 2015), for which state-of-the-art speech recognition systems reach a performance of 1.4% WER (Zhang et al., 2020). Also the read speech components of the Kiel (Kohler et al., 2017) and GRASS corpus (Schuppler, Hagmüller, et al., 2014) used in this study fall into this category (cf. Sec. 2.2 for more detail).

Next, we distinguish spontaneous speech with respect to whether it is task-oriented or not. Speech from task-oriented dialogues are in general characterized by

---

This chapter has been reformatted from:

[A] Julian Linke, Saskia Wepner, Gernot Kubin, and Barbara Schuppler. (2023). Using Kaldi for automatic speech recognition of conversational Austrian German. *ArXiv* (abs/2301.06475).

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

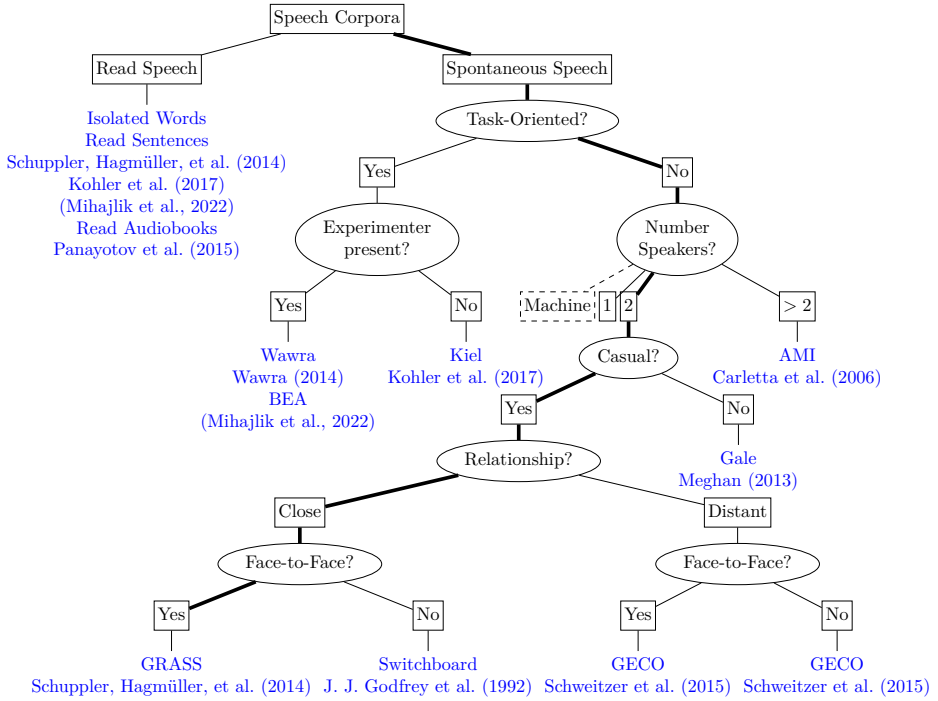


Figure 2.1: Categorization of speech corpora for different speaking styles. The tree structure is defined by statements (black rectangles), questions (ellipses) and chosen examples (blue). The dashed rectangle indicates the possibility of machine-oriented interaction between speakers and machines (i.e., dialogue systems). The bold line indicates the path for GRCS.

covering a specific domain, and that speakers chat less freely than when the topic of a conversation is open. Task-oriented dialogues can be categorized based on the presence or absence of an experimenter (such as a linguist or broadcast interviewer) who guides the conversation. This type of speech is characterized by utterances of relatively complete syntactic structures which are pronounced carefully, given that trained speakers are involved. In contrast, in casual conversations, speakers strongly reduce their pronunciation and produce syntactically incomplete structures (Johnson, 2004). An example for task-oriented dialogues without experimenter present are the dialogues in Verbmobil (Wahlster, 1993) and in the Kiel Corpus (Kohler et al., 2017). These short dialogues last for approx. $2 - 20$ minutes each, a time span that does not allow the speakers to *forget about the recording situation*, which affects the naturalness of the resulting speaking style.

Another way how to categorize speech corpora is with respect to the (number of) interlocutor(s). We distinguish (spontaneous) monologues and machine-oriented dialogues that both do not show cross-talk, from conversations between two or more speakers (e.g., conversations between three speakers in a casual setting (Torreira et al., 2010); between even more speakers in a meeting setting (Carletta et al., 2006)). With increasing number of speakers, the challenge for ASR is rising, as one needs to deal with overlapping speech, which comes not only with acoustic difficulties, but also with structural speech phenomena such as co-completion, turn-competition and broken turns. For the AMI-Meeting corpus, WERs of approx. 21.2% have been

achieved (Kanda et al., 2021).

When focusing on conversations between two speakers, we may further distinguish whether the topic of the conversation is casual or professional, as we assume that casual topics also lead to a speaking style that is characterized by more pronunciation variation and/or a stronger use of dialectal variants. Pronunciation variation may become even more salient when speakers have a close relationship to each other (and are maybe even from the same dialectal area). All of these effects on style may be continuous in some language areas (e.g., in Austrian German), or diglossic in others (e.g., in Swiss German) (Stępkowska, 2012). We are aware of the sociolinguistic fact that the effect of relationship on the speaking style is not comparable across languages.

Finally, we categorize conversational speech (CS) with respect to whether they occurred face-to-face or not. In the widely used Switchboard corpus (J. J. Godfrey et al., 1992), speakers who knew each other well were having telephone conversations, where the speakers were not able to benefit from visual cues and needed to deal with reduced sound quality, forcing them to pronounce more clearly and to avoid overlapping talk. ASR results for Switchboard are in the range of 4.3 % to 5.1 % WER (Tüske et al., 2021; Xiong et al., 2018).

The IMS GECO database from Stuttgart contains conversations between two speakers of a distant relationship, in face-to-face (GECO-Multi) and in a non-face-to-face setting where speakers were separated by a solid wall (GECO-Mono) (Schweitzer & Lewandowski, 2013; Schweitzer et al., 2015). First word recognition results with GECO correctly identified only 25 % of the words (Arnold et al., 2017).

The corpus in focus of this thesis is the conversational component of the GRASS corpus (GRCS), containing topic-open, casual, face-to-face conversations between two closely related persons that last for one full hour, with no experimenter present (Schuppler et al., 2017). So far, there is little data available for this specific speaking style. The Japanese *Corpus of Everyday Japanese Conversations* (CEJC) (Koiso et al., 2018), published in March 2022 (Koiso et al., 2022), includes recordings collected through an individual-based recording method. The material comprises recordings from 40 informants balanced in terms of sex and age, each collecting approx. 15h of speech data using portable recording devices over two to three months in various everyday situations. To the best of our knowledge, so far there have not yet been published any ASR experiments with CEJC; a study on dialogue situation recognition using CEJC showed that the system did not reach the level of human evaluation results (Chiba & Higashinaka, 2021). In summary, CEJC and the data used in this study (GRCS) contain a broad variety of challenges resulting from speaker interaction in conversational speech.

2.2 Spontaneous speech corpora

This thesis presents experiments on spontaneous speech corpora, with a primary focus on Austrian German conversational speech (cf. Sec. 2.2.1). To explore regional variations, some experiments also include other German corpora, namely the IMS GECO database (cf. Sec. 2.2.2) and the Kiel Corpus (cf. Sec. 2.2.3). In addition, one experiment extends the analysis by including the Hungarian BEA database (cf. Sec. 2.2.4), which provides additional insights into cross-lingual speech representations (cf. Sec. 4.4).

2.2.1 GRASS corpus

The Graz corpus of Read And Spontaneous Speech (GRASS) (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017) contains about 30h of Austrian German read (GRRS) and conversational speech (GRCS) from 38 Austrian speakers (19f/19m). As language usage in CS varies strongly with educational level, social background and dialect region, speakers were selected who were born in the same broad dialect region (Eastern Austria), had been living in an urban area for years and had a higher education degree. For the CS component, 19 pairs of speakers who have been knowing each other for several years were recorded for one hour each without interruption in order to encourage a fluent, spontaneous conversation. There was no experimenter present in the recording room and there was no restriction in terms of chosen topic or speaking behavior, leading to the use of natural, partly dialectal pronunciation with typical characteristics such as frequently occurring overlapping speech, laughter, and the use of swear words (Schuppler et al., 2017). Despite the speakers' awareness of being recorded, they appeared to completely forget about the studio recording situation after a period of five to ten minutes, entering a casual conversation. Only after the hour of CS, speakers read short stories as well as selected isolated sentences. Both, RS and CS component were produced by the same speakers.

2.2.2 IMS GECO database

The IMS GECO database (GECO) (Schweitzer & Lewandowski, 2013; Schweitzer et al., 2015) contains 46 spontaneous dialogues of approx. 25 minutes between unfamiliar female speakers from two settings: 1) a unimodal setting with 22 dialogues (GECO-Mono; GEMO), where participants could not see each other because they were separated by a solid wall and 2) a multimodal setting with 24 dialogues (GECO-Multi; GEMU), with face-to-face conversations comparable to GRCS. The unimodal setting involves 12 speakers, where 7 returned for the multimodal setting meaning that some dialogue pairs are present in both GECO-Multi and GECO-Mono. In both settings, speakers were free to choose the topics they wanted to discuss.

2.2.3 Kiel corpus

The Kiel Corpus of Spoken German (Kohler et al., 2017) contains a total of 5 h of read (KIRS) and spontaneous speech produced by speakers mainly coming from Northern Germany. Two spontaneous components are available: (1) the "appointment-making-scenario" part (KIVM), which contains approx. 4 h of dialogues from 43 speakers (22f/21m) who were making appointments. In this scenario, speech was only recorded if participants were holding a button pressed which was at the same time blocking the interlocutor's channel. Thus, this scenario effectively avoids overlapping speech. (2) the "video-task-scenario" part (KIVT) contains approx. 1 h of dyadic conversations. In this scenario, manipulated video materials from a television series were presented separately to two participants who had the task to find the differences in the video. We used the spontaneous speech component from the Kiel Corpus (KICS) for our experiments with Austrian CS.

2.2.4 BEA database:

The original BEA ("BEszélt nyelvi Adatbázis" in Hungarian, meaning spoken language database) aimed at collecting studio quality speech data from 500 speakers, representative in age, sex, dialect, and educational background, primarily for linguistic research purposes (Neuberger et al., 2014). The BEA-Base subset (Mihajlik et al., 2022) of the database includes the read *Readtext* (BERS) and the conversational *Discourse* (BECS) modules of the "train-114" subset. Both, BERS and BECS included the same speakers while female and male participants were closely balanced. In case of BECS, each conversation was recorded approx. 45 min and one experimenter guided the casual conversations between the speaker and an optional discourse partner on various random topics. The recordings were made in the same studio environment and were cleaned from ambiguous and parallel parts, similarly to the previous databases. The recordings containing the voices of the experiment leader or of a 3rd person were excluded from the investigations. Hence, conversations from BECS included recordings which relate to only one speaker which makes it possible to compare specific speakers between BECS and BERS but, different from GRCS, it is impossible to compare one speaker pair from BECS with respective speakers from BERS.

2.3 Initial Kaldi experiments for Austrian German

This section introduces initial Kaldi experiments for Austrian German read and conversational speech. These initial speech recognition experiments reveal that the conversational speaking style poses challenges, especially with respect to the GRASS corpus which supports a low-resourced language processing assumption (cf. Sec. 4.3).

2.3.1 ASR for read speech

Methods: Acoustic Models (AM) and Language Models (LM) were trained on data from the RS corpus. The GRRS data set comprises 6h of speech, where each speaker reads mostly the same, phonetically balanced sentences. The training set included 33 speakers (5.25h), the validation set 2 speakers (0.37h) and the test set also 2 speakers (0.34h).

We extracted 13-dimensional MFCCs and performed cepstral mean and variance normalization (CMVN) while comparing a combination of different frame lengths $\{20\,\text{ms},\ 25\,\text{ms},\ 30\,\text{ms}\}$ and frame shifts $\{7.5\,\text{ms}, 10\,\text{ms}, 12.5\,\text{ms}\}$. For the acoustic models (AM), the initial diagonal GMM-HMM models (short GMM) comprise basic monophone and triphone training with MFCCs+Δ+$\Delta\Delta$ features.

The lexicon was built with a G2P online tool (Reichel, 2012) for standard German. As this resource is not available for the Austrian variety of German, we applied a set of rules on phone-level to adapt its output towards standard Austrian German pronunciation (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014), and phonological reduction phenomena, such as schwa-deletion. We reduced the phone set yielded by G2P in order to improve recognition performance using three rules: (R1), a replacement rule to devoice all alveolar and postalveolar fricatives and affricates (a common phonological process in standard Austrian German); (R2), a

Table 2.1: Summary of the best WERs [%] with GRRS. We trained a GMM (cf. Sec. 2.3.1) and compared the impact of different phone set minimization rules on *valid* and *test* sets.

Phone Set Rule	*valid*	*test*
–	0.96	1.2
+R1	1.05	1.04
+R1+R2	**0.4**	**0.64**
+R1+R2+R3	0.56	**0.64**

rule to split all diphthongs into two separate phones; (R3), a rule to unite short and long vowels, based on phonetic studies on Austrian German (Moosmüller, 2007). In total, we reduced the phone set from initially 64 to 38 phones.

We used the SRILM toolkit with a Witten-Bell discounting for an N-gram language model (LM) of different orders (Stolcke, 2002). Note that the LM for the GRRS experiments was generated given the text of all utterances from the entire RS component since all speakers in the RS component read the same text[1].

Results: Tab. 2.1 shows a summary of the results for RS. First, we analyzed the influence of different frame shifts and frame lengths with a trigram LM. Our experiments showed that different frame lengths of $f_{\text{len}} = \{20\,\text{ms}, 25\,\text{ms}, 30\,\text{ms}\}$ have less impact on the WERs than frame shifts $f_{\text{sh}} = \{7.5\,\text{ms}, 10\,\text{ms}, 12.5\,\text{ms}\}$. The best triphone WER (0.96 %) was achieved with $f_{\text{sh}} = 12.5\,\text{ms}$ and $f_{\text{len}} = 20\,\text{ms}$. Monophone and triphone models performed similarly most of the time. Yet, with a frame shift of $f_{\text{sh}} = 10\,\text{ms}$ combined with frame lengths of $f_{\text{len}} = \{25\,\text{ms}, 30\,\text{ms}\}$, triphone models returned worse results. In order to further optimize the set of AMs, rules R1, R2 and R3 were applied one after the other. When comparing our final WERs with the RS component (cf. Tab. 2.1), R1 and R2 lead to an improvement of our best triphone WERs by 0.56 %. R3 slightly deteriorated our results by 0.16 %.

Next, we tested different LM of orders $\{1, 2, 3, 4, 5\}$ with our best frame shift $f_{\text{sh}} = 12.5\,\text{ms}$ and frame length $f_{\text{len}} = 20\,\text{ms}$ configuration. Bigrams, trigrams, four-grams and five-grams performed similarly well regarding both, monophone and triphone models (WER $\approx 1\% - 2\%$). We decided to stick to the trigram model which had a slight advantage with the best WER = 0.96 %. With unigrams, we achieved non-comparable results since best WERs differed widely. In this case, best WERs of the monophone model (WER = 35.77 %) were also much worse than best WERs of the triphone model (WER = 18.25 %).

Conclusion: Our ASR experiments for RS showed that the lowest WER was achieved with a lexicon with canonical pronunciation, i.e. no pronunciation variants. The only adaptation made to this canonical lexicon was the reduction of the phone set according to Austrian Standard German pronunciation (e.g., devoicing alveolar fricatives). State-of-the-art performance was obtained with a basic triphone model (0.64% WER with *test*). We observed that our methodological choices lead to large improvements (i.e., frame shift, phone set minimization, AM passes and LM orders). Additionally, the difference in WER between *valid* and *test* were relatively

[1] In Sec. 4.2.4, we describe GRRS experiments based on different training splits also with respect to the LMs that achieved a best speaker-dependent WER of 0.67%.

low (approx. $0.01\% - 0.24\%$). In general, the WERs were in the range of other state-of-the art systems for RS (e.g., 1.4% in Chung et al. (2021)).

2.3.2 ASR for conversational speech

Methods: Fig. 2.2 shows a schematic overview of the experimental setup. AMs and LMs were trained with data from GRCS, GECO and KICS (cf. Sec. 2.2). We present experiments which are trained merely with GRCS, or GRCS and GECO, or GRCS and KICS or GRCS and GECO and KICS. This study focuses on evaluating ASR on conversational Austrian German by performing leave-p-out cross-validation with respect to GRCS (with $p = 2$ speakers of the same conversation) resulting in approx. 0.8 h of test data and 13.5 h of training data per split. We randomly chose 10% of resulting training splits as validation sets (approx. 1.35 h) to adjust basic model parameters. When adding training data from GECO or KICS, validation sets were built by randomly choosing 10% from the newly introduced training data. For evaluation, we compared the performance on the test splits which result from the described cross-validation.

In GRCS preprocessing, we excluded chunks that contained laughter, singing, imitations/onomatopoeia, completely incomprehensible word tokens or artefacts (e.g., when a speaker accidentally touched their microphone). In case of GECO, we removed symbols indicating laughter, throat clearing and broken words from the transcriptions. In case of KICS, we removed symbols indicating laughter, smacking sounds, different types of noise and repetitions from the transcriptions.

For the AM, the ASR monophone and triphone training steps were in most parts analogous to the RS experiments. First two models were again trained with 13-dimensional MFCCs+Δ+$\Delta\Delta$ and CMVN (with $f_{\text{sh}} = 10\,\text{ms}$ and $f_{\text{len}} = 25\,\text{ms}$). On top of the triphone GMMs (cf. Sec. 2.3.1), a speaker independent GMM model with linear discriminative analysis (LDA) and maximum likelihood linear transform (MLLT) (Gopinath, 1998) was trained resulting in GMM+LDA+MLLT. Speaker-adapted training was performed also on top of GMM+LDA+MLLT with constrained maximum likelihood linear regression (fMLLR) (Gales, 1998) resulting in GMM+fMLLR. The final triphone alignments were used to train a baseline DNN-HMM hybrid model consisting of a TDNN with 13 layers and hidden dimensions of 512 while utilizing only already calculated MFCC features. The network is trained with a frame-level objective function based on the cross-entropy criterion. Our recipe is based on a recipe published in Meyer (2020) and related DNN setups are described in Rath et al. (2013) and Veselý et al. (2011).

For the LM, we used the SRILM toolkit (Stolcke, 2002) with the same configuration as in RS but trained trigrams. Here, the LM was generated given the text of all utterances from the cross-validation training splits from GRCS and the two additional German corpora if they were also utilized for AM training. In order to evaluate a potential limited data problem when training the LMs, we also ran experiments utilizing a bigger trigram by adding approx. 220 k Austrian German sentences from subtitles of broadcasts for the deaf and hard of hearing of an Austrian public television service (AGS) (*ORF-TVthek: Broadcasts for the Deaf and Hard of Hearing*, n.d.). For the latter, we performed LM-rescoring with a four-gram by again adding a random subset of 5M German sentences from AGS, German Wikipedia[2]

[2] `https://dumps.wikimedia.org/dewiki/20220701/dewiki-20220701-pages-articles`

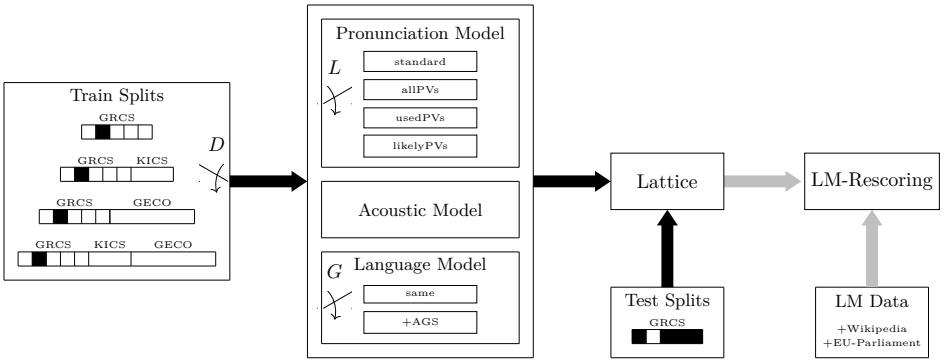


Figure 2.2: Schematic architecture for our conversational speech recognition experiments. Training is conditioned on 4×19 data combinations since we introduce 4 possible corpora combinations (switch D) and each conversation from GRCS is processed individually. This results in 19 conversation-dependent training and test splits for 4 data combinations (cf. Tab. 2.3). Experiments analyzing the influence of the lexicon (switch L) utilize the same data for AM and LM training. Experiments analyzing the influence of the LM (switch G) utilize our standard lexicon and perform LM-rescoring as an additional option (grey arrows).

and the European parliament[3]. In order to receive the additional LM data, we adapted a toolkit which is described in Milde and Köhn (2018).

Lexicon Generation: We created word lists from the transcriptions of all corpora. Canonical pronunciations were obtained with a G2P online tool (Reichel & Kisler, 2014). We created four different pronunciation lexicons.

standard. Since the German language setting of the utilized tool creates pronunciations for German Standard German (GSG)[4], we applied 6 input switch rules to obtain an Austrian Standard German pronunciation. We call the resulting pronunciation lexicon *standard*. For the foreign language words, we changed the language setting to the corresponding language.

allPVs. We created a lexicon with pronunciation variants (PVs) by applying 26 phonological rules (based on findings from Schuppler, Adda-Decker, and Morales-Cordovilla (2014)) to the canonical Austrian German pronunciations. 17 of these rules are relevant for conversational speech of all German varieties, e.g., assimilation of plosives and r deletion in the syllable coda, whereas 9 rules cover pronunciations that are typical for the Austrian German variety, e.g., the deletion of the syllable "ge" in the beginning of specific past participles. In addition to these rule-based variants, we also created a couple of PVs manually, in order to capture pronunciations that cannot be generated in an automated way but are frequent in Austrian German spontaneous speech, e.g., the pronunciation [maːɪ] for the word "wir" with canonical pronunciation [viːɒ]. The resulting lexicon (*allPVs*) contained on average $5.57 - 6.18$ variants per word.

usedPVs. We used the *allPVs* lexicon to create a forced alignment based segmentation (with a frame shift of $f_{\text{sh}} = 7.5\,\text{ms}$) of all of corpora (i.e., GRCS, KICS and GECO). From these segmentations, we extracted the pronunciation variants that

-multistream.xml.bz2

[3] https://www.statmt.org/europarl/v7/de-en.tgz

[4] With German Standard German, we refer to German as spoken by speakers from Germany

Table 2.2: Entry statistics of the different pronunciation lexicons used for the ASR experiments with conversational speech. The number of entries is influenced by the number of utilized corpora for AM/LM training (cf. Tab. 2.3 and Fig. 2.2).

Lexicon Name	min(**#entries**)	max(**#entries**)
standard	13.9k	22.7k
allPVs	74.2k	135.1k
usedPVs	17.4k	30.4k
likelyPVs	14.6k	26.9k

had been actually produced by the speakers, and created a pronunciation lexicon with those PVs only. The resulting lexicon (*usedPVs*) had on average $1.37 - 1.43$ variants per word.

likelyPVs. We created a lexicon containing only those variants which showed a high frequency of occurrence in the forced alignment, inspired by the approach presented in Chen et al. (2015). As in Chen et al. (2015), we calculated the statistics for the pronunciation probability estimation, but instead of integrating different probabilities for specific pronunciations, we considered only pronunciations which result in an estimated probability of $p > 0.65$. This choice was made in order to give a better comparison with our other lexicons since introducing additional pronunciation probabilities to the lexicon would change the experimental design (in other words, all final lexicons involve pronunciation variants with equal probabilities by definition). The resulting lexicon (*likelyPVs*) had an average of $1.16-1.26$ variants per word. Tab. 2.2 presents a summary of all different pronunciation lexicons used in our ASR experiments.

Results: Tab. 2.3 shows the ASR results for different training setups, always using GRCS as test data. We compared ASR experiments with (1) different data sizes for AM training, (2) lexicons of different amounts of variants (3) LMs trained on different data sizes and (4) a combination of the best AM, Lexicon and LM. With respect to the Acoustic Model (AM), all AMs showed benefits from additional training data from other corpora, resulting in absolute WER improvements of approx. $1\% - 2\%$ with respect to mean values; then again, respective standard deviations are higher when more data is used indicating that overall performance improves but robustness problems arise. With respect to the pronunciation lexicon, our results showed that in comparison to using our standard lexicon, lexicons with very high numbers of variants (i.e., *allPVs* and *usedPVs*) lead to a performance decrease. With the *likelyPVs* lexicon, however, which contained a small number of likely pronunciation variants, performance improved by approx. 1.5% compared to the best mean value of the baseline with the standard lexicon. With respect to varying the amount of training data for the LM, we achieved the best results by adding data from all corpora for AM training, adding data from all corpora plus AGS for LM training, a lexicon with likely pronunciation variants and LM-rescoring with our 5M additional German sentences (cf. Sec. 2.2) resulting in a best mean WER of 48.5%.

Overall, when comparing the best mean WER with our baseline system, we achieved an absolute WER improvement of approx. 4.5%. In general, in all experi-

Table 2.3: Summary of conversation-dependent WERs [%] for Austrian German conversational speech obtained with our Kaldi baseline system. The first two columns show the utilized data for AM and LM training, the third column shows the utilized lexicons and the remaining columns give mean and standard deviations of resulting 19 WERs as well as corresponding minimum (min) and maximum (max) WERs.

	AM	**LM**	**Lexicon**	**WERs**	**min**	**max**
Baseline	GRCS	same	standard	53.89/5.18	42.1	63.7
	GRCS+KICS			53.22/5.23	42.0	63.2
	GRCS+GECO			52.58/5.59	40.5	63.9
	GRCS+GECO+KICS			**51.91/5.74**	**39.7**	**63.2**
Influence Lexicon	GRCS	same	**allPVs**	55.56/5.03	43.9	64.7
	GRCS+KICS			55.16/5.21	44.4	64.8
	GRCS+GECO			54.37/5.45	42.6	64.7
	GRCS+GECO+KICS			53.62/5.56	41.6	64.7
	GRCS	same	**usedPVs**	55.22/4.88	43.8	64.3
	GRCS+KICS			55.06/5.15	43.5	64.5
	GRCS+GECO			54.15/5.4	42.5	64.7
	GRCS+GECO+KICS			53.69/5.52	42.9	64.3
	GRCS	same	**likelyPVs**	51.87/4.88	40.3	**60.9**
	GRCS+KICS			51.64/5.22	40.0	61.9
	GRCS+GECO			50.93/5.61	38.8	62.2
	GRCS+GECO+KICS			**50.48/5.66**	**39.7**	62.0
Influence LM	GRCS	**+AGS (220k)**	standard	52.26/5.5	40.1	62.9
	GRCS+KICS			51.53/5.58	40.3	62.8
	GRCS+GECO			51.38/5.72	40.3	63.3
	GRCS+GECO+KICS			50.74/5.82	39.0	62.3
	GRCS	**+AGS (220k) +Rescoring (5M)**	standard	51.17/5.26	39.5	62.4
	GRCS+KICS			50.91/5.65	39.8	62.4
	GRCS+GECO			50.92/5.82	39.5	63.2
	GRCS+GECO+KICS			**50.19/5.82**	**38.7**	**62.2**
Best	GRCS	**+AGS (220k) +Rescoring (5M)**	**likelyPVs**	49.15/5.28	37.3	**59.1**
	GRCS+KICS			49.02/5.69	37.3	60.4
	GRCS+GECO			49.07/5.88	37.2	60.8
	GRCS+GECO+KICS			**48.5/6.09**	**37.0**	61.3

ments, we observed highly varying WERs between the different conversations (i.e., speaker pairs) of GRCS (standard deviations range from 4.88% to 6.09%).

Discussion: These experiments aimed at building a Kaldi-based ASR system for Austrian German, with a focus on conversational speech. Since our first experiments already showed large differences from speaker pair to speaker pair, we decided to provide cross-validation results in order to get more insight into conversation-dependency of ASR systems. It is worth noting that, even though when reaching

performance gains by certain methodological choices, we still observed similarly high standard deviations of the WERs across the conversations. Hence, neither the change of data sizes for AM and LM training nor the different approaches for pronunciation modeling made the ASR system more robust to variation in conversational speech (Linke et al., 2022).

In comparison to other benchmarks, our results form the cross-validation approach highlights how challenging the task of conversational speech recognition is. Other benchmarks tend to train and test on pre-defined training and test sets which may cause an optimistic bias towards ASR accuracy (Szymański et al., 2020). The cross-dialect analysis described in (Elfeky et al., 2018), for instance, showed how ASR performance decreases when a dialect variation is evaluated on a system which had been trained on another dialect of the same language. We hypothesize that testing each conversation individually shows a similar effect because even though speakers in GRASS had a comparable regional background, we still find high individual dialectical variation which is in line with a previous analysis of the corpus in (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014). Our results suggest not only to investigate how to improve overall ASR performance but to focus more of tackling missing robustness, especially in case of conversational speech recognition.

ASR with our standard lexicons and a large amount of additional LM data resulted in a mean WER of 50.19%. When utilizing a lexicon with likely pronunciation variants only (by adding approx. 4.2k entries to the standard lexicon; cf. Tab. 2.2) without adding a large amount of additional LM data, we achieved a mean WER of 50.48%. Thus, when comparing our results from using different lexicons with those from using different amounts of data for the LM (cf. Tab. 2.3), we observe that training LMs with large amounts of data had a similarly high impact on improving WERs (approx. 1.5%) as using the best pronunciation lexicon (i.e., *likelyPVs*). A survey on modeling pronunciation variation for ASR (Strik & Cucchiarini, 1999) summarizes that adding pronunciation variants to the lexicon appears to improve recognition performance especially if the different frequencies of occurrence of variants are considered. Two decades and many ASR architectures later, our results still confirm their observation. We further showed that well-developed pronunciation modeling, for which no additional data resources nor high computational efforts are necessary, could compensate for the necessity of collecting more LM data. Yet, the combination of both methodological approaches (pronunciation modeling and collecting more LM data) still yielded the best results. This finding is especially relevant for the field of low-resourced ASR.

From the literature, we know that expanding the lexicon to include pronunciation variants can improve coverage of observed variation in spontaneous speaking styles, but it can also increase the search space, which may lead to higher decoding complexity and degraded recognition performance (Karanasou, 2013; Kessens et al., 2002; Strik & Cucchiarini, 1999). We analyzed the search space of the ASR decoders for the different pronunciation lexica (*standard*, *allPVs*, *usedPVs* versus *likelyPVs*), focusing on the interaction between mean lattice depth per frame, mean WERs and pronunciation variants per word. Mean lattice depth, calculated during the beam search decoding process in Kaldi, represents the average number of competing hypotheses (arcs) crossing each frame in the search lattice. This metric directly reflects the complexity of the decoder's search space, as more pronunciation variants lead to more competing paths that must be maintained within the beam.

For systems trained exclusively on GRCS, our analysis showed that the system utilizing likely pronunciation variants (1.16 variants per word) achieves the best balance with a mean WER of 51.87% and the lowest mean lattice depth of 71.58. In comparison, the baseline system with standard pronunciations (1 variant per word) shows moderate performance with a mean WER of 53.89% and mean lattice depth of 110.39, while systems with more variants (i.e., *utilizedPVs* and *allPVs*) demonstrate performance degradation. Specifically, increasing to 1.37 variants per word leads to higher search complexity (mean lattice depth of 125.89) and worse recognition performance (mean WER of 55.22%). The trend continues with the inclusion of all pronunciation variants (5.57 variants per word), resulting in the highest mean lattice depth of 138.19 and poorest performance with a mean WER of 55.56%. These results align with previous research emphasizing the importance of balancing pronunciation variant modeling in ASR. While some degree of variation improves robustness to real-world speech phenomena, excessive numbers of variants can inflate search space and lead to increased complexity in the decoder. Our findings reinforce the principle that lexicon design should prioritize quality over quantity by focusing on likely pronunciation variants, ensuring an optimal balance between recognition accuracy and search efficiency.

Conclusion We achieved similar performance gains by either incorporating knowledge into the pronunciation lexicon or augmenting the training data. We observed high variation in performance from conversation to conversation (i.e., approx. $5\% - 6\%$ standard deviation), regardless of the overall performance, indicating low robustness of the ASR system for conversational speech. The reasons for the lack of robustness could come from high variation with respect to pronunciation variation (i.e., dialectal background), speech rate variation (in CS speech rate varies from 0.88 to 45.45 phones per second with a mean of $12.38\,\mathrm{s}^{-1}$ and a standard deviation of $4.28\,\mathrm{s}^{-1}$), differences in lexical choice (as the topics are chosen freely), differences with respect to whether complete syntactic structures are used by the speakers and their turn-taking behaviour. In future work, we plan to analyze in detail which are the factors that hinder robust ASR of conversational speech.

Chapter 3

Prosodic prominence

3.1 Introduction

Prosodic prominence is a complex phenomenon that can be studied from different perspectives (B. Wagner et al., 2015). Some linguistic unit is usually considered prosodically prominent if it is perceived as standing out from its environment (Terken, 1991). However, what is perceived as prominent is influenced by a multitude of factors involving functional, structural and frequency criteria, and the rating task as much as the physical properties of the item that is perceived as being prominent (Bishop, 2012; Cole et al., 2019, 2010; Turnbull et al., 2017; P. Wagner, 2005). In the following works, we focus on the correlates of prominence in the acoustic signal.

3.1.1 Acoustic cues and perception of prosodic prominence

Different languages rely on different weightings of acoustic cues to create perceptual prominence (Beckman, 1986), the most important of which are F0 variation, duration and intensity (e.g., (Baumann et al., 2016; Kochanski et al., 2005; Mixdorff et al., 2015; Terken & Hermes, 2000; Turk & Sawusch, 1996)). With respect to F0, not only pitch height or excursion are relevant for prominence, but also the shape and alignment of pitch contours (Baumann et al., 2016; Kohler & Gartenberg, 1991; Niebuhr, 2010). For English, Cole et al. (Cole et al., 2010) found that duration is a more important cue for prominence perception than intensity/loudness (e.g., (Arnold et al., 2012; Turk & Sawusch, 1996)).

For German, Arnold et al. (2013) and Niebuhr and Winkler (2017) reported that F0 was a more important correlate of prosodic prominence than syllable duration. Pitch-accent related variables outranked both acoustic F0 and durational features in Baumann and Winter (2018). On the other hand, Arnold et al. (2012) argued that word duration was more important for prominence perception than F0 and intensity,

---

This introduction has been reformatted from:

[B] Julian Linke, Anneliese Kelterer, Markus A. Dabrowski, Dina El Zarka, and Barbara Schuppler. (2020). Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proc. of Speech Prosody* (pp. 1000–1004).

[C] Julian Linke, Gernot Kubin, and Barbara Schuppler. (2023). Using word-level features for prosodic prominence detection in conversational speech. In *Proc. of ICPhS* (pp. 3101–3105).

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

and Tamburini and Wagner (2007) found that force-accent related parameters (i.e., duration and spectral emphasis) were more important for syllable prominence than pitch-accent related parameters (i.e., F0 contour features and overall intensity). For Austrian German, it has been found that duration and spectral tilt are strong acoustic cues to perceptual prominence, whereas a change in F0 within a syllable did not necessarily correlate with stronger perceptual prominence (El Zarka et al., 2017). Concerning vowel quality, higher prominence only seems to affect F1, but not F2 and F3 (El Zarka et al., 2017).

3.1.2 Automatic prosodic annotation tools

Several automatic prosodic annotation tools have been built and distributed. Some of them combine acoustic, lexical and syntactic features (e.g., Ananthakrishnan and Narayanan (2008) for American English, Avanzi et al. (2008) and Christodoulides et al. (2017) for French), others use lexical and syntactic information alone (e.g., Marsi et al. (2003) for Dutch). Arnold et al. (2013) used GAMs and random forests to model prosodic prominence in German, with the aim of analyzing and comparing the contribution of acoustic, linguistic and contextual information. Like Arnold et al. (2013), we aim at using random forest models to learn more about the contribution of the features to prosodic prominence perception. Since we additionally aim at building a tool that can be incorporated into the annotation process of a not-yet annotated database, a requirement for the tool is the use of acoustic features alone.

For American English, Tamburini and Caini (2004) proposed a tool that classifies whether a syllable is prominent or not. The prediction was based on the speech waveforms only, with no higher level linguistic information available to the tool. For German, only a few prosodic annotation tools have been built that use acoustic features alone. Braunschweiler (2003), for instance, proposed ProsAlign, a system that automatically produces GToBI labels. The tool covers 56% of the manually established labels and can thus be integrated in a semi-automatic annotation procedure. Since the development of ProsAlign, however, other prosodic annotation systems than GToBI have been developed for German (e.g., KIM, DIMA) (Kügler et al., 2019). The tool by Tamburini and Wagner (2007) annotates prominence as a continuous, rather than a categorical parameter. Their analysis led to the conclusion that force accents are a more reliable cue to prominence than pitch accents in German. For Austrian German, no tool is available at this point.

3.1.3 Entropy-based prosodic features

So far, most studies on prosodic prominence analyzing F0/RMS contours considered features related to specific characteristics of those curves (i.e., mean, maximum, etc.). To the best of our knowledge, it has not yet been investigated whether entropy-based F0/RMS features, which directly relate to their distribution distinguishing prominence levels. In general, entropy-based features have broadly been used in speech science: For instance, Klabbers and Veldhuis (2001) use relative entropy to measure the distance between two speech spectral distributions in concatenative synthesis applications whereas Misra et al. (2004) showed that spectral entropy features which interpret the spectrum as a probability mass function, improved the performance of an automatic speech recognition (ASR) system. A study on voice signal characterization tested entropy measures coming from raw audio signals in

order to extend voice analysis methods (Rogério Scalassara. et al., 2008). With respect to prosody of emotional expressions, it has been shown that the use of features capturing F0/RMS variability by calculating entropy from F0/RMS contours helps to distinguish between arousal conditions in a free-speech setting (Cohen et al., 2010).

3.2 Prominence classification in read speech

3.2.1 Materials and methods

GRASS corpus: This study is based on read speech from GRASS (GRRS) (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017). GRRS comes with automatically created segmentations using MAUS (Kisler et al., 2017), which were corrected manually. GRRS was manually annotated prosodically, using the same criteria as the Kiel corpus (IPDS, 1997), with prominence ratings 0 (no prominence; *PL0*), 1 (weak prominence; *PL1*), 2 (strong prominence) and 3 (emphatic prominence). In this study, we combined prominence levels 2 and 3, as empathic prominence occured rarely (*PL2*). Three phonetically trained transcribers created the prosodic annotations in the following way: one transcriber created the first version of the annotation, which was subsequently corrected by the other transcribers. This procedure reached a high inter-annotator agreement (Cohen's kappa: 0.81, 0.76, 0.63, calculated on 269 word tokens from 47 utterances). Prominence classification results in Sec. 3.2.2 are based on a training set of 197 utterances (2919 word tokens) from two narratives, read by 10 male and 9 female speakers. The test set for the classification experiments consists of 47 utterances (269 word tokens) annotated by all three annotators and the prominence ratings for the test set were assigned by majority decision. In contrast to the training set, the test set primarily consisted of isolated short sentences ($\approx 80\%$ of the utterances in the test set included 5.3 ± 0.8 word tokens in comparison to the training set with 15.1 ± 7.4 word tokens per utterance).

Acoustic feature extraction: For each word, we extracted 96 features based on the fundamental frequency F0, the sound intensity (RMS) and durational characteristics. F0 was calculated with the library *AMFM decompy* (Schmitt, 2018). This package contains an implementation of the pitch detection algorithm *YAAPT* (Zahorian & Hu, 2008). Sound intensity was calculated directly from the waveform by calculating the root mean square. For F0 and RMS, and their respective first and second derivatives, 10 measurements were extracted: maximum, minimum, range, relative position of maximum and minimum in the word, mean, median, first and third quartile and standard deviation (60 features). For the basic F0 and RMS curves, we extracted 12 measurements: left and right slope of the maximum and minimum, absolute and relative onset and offset within the word, as well as maximum, minimum, range and mean relative to the utterance (24 features). We employed the peak detection algorithm (Duarte & Watanabe, 2018) and *numpy* (Walt et al., 2011) for the statistial features. The 12 durational features were: word duration, total speech rate (phrase), local speech rate (word), and 9 relative speech rate measures. The local speech rate is estimated as the ratio of number of segments to word length.

This section has been reformatted from:

[B] Julian Linke, Anneliese Kelterer, Markus A. Dabrowski, Dina El Zarka, and Barbara Schuppler. (2020). Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proc. of Speech Prosody* (pp. 1000–1004).

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

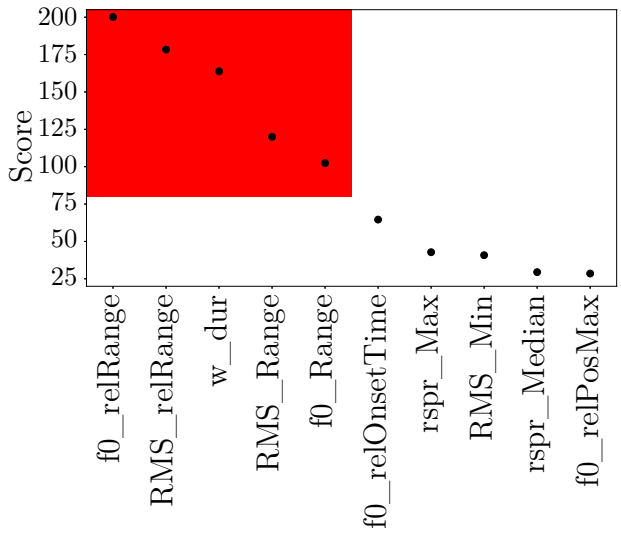


Figure 3.1: 10 highest χ^2-scores in case of a classification task with 3 classes. The selected features are located in the red marked area. A selection of the first 3 features did not result in a satisfying performance. Adding 2 more features indicated a more reliable classification performance, which can be compared to the ranking of a RFC with all 96 features included (cf. Sec. 3.2.2).

Classification methods and testing: Two classification methods were implemented in Python with the *scikit learn* toolkit (version 0.21.3.) (Pedregosa et al., 2011). First, we modelled a decision tree (DT) with only two classes and the Gini Impurity (I_G) as an impurity measure. For the purpose of this classification task, prominence ratings 1, 2 and 3 were combined in the class *PL12* (prominent, 1669 tokens). The associated null class is called *PL0* (non-prominent, 1250 tokens). This binary classification task has the aim to test how well prominence classification is at all possible with the given data set compared to prior studies. Parameterization was done by comparing the results of different DT-models with varying tree depths fitted with the training set. A comparison of different DT-topologies showed that a depth of 1, which obviously leads to a highly simplified model, was sufficient to distinguish between the classes *PL0* and *PL12*. Hence, including higher depths resulted in more complicated models with no improvement in the respective F1-scores.

Second, a Random Forest classifier (RFC) with 1000 decision trees was used for a classification task with three classes: *PL0* (no prominence, 1250 tokens), *PL1* (weak prominence, 726 tokens), and *PL2* (strong prominence, 943 tokens; prominence ratings 2 and 3 were subsumed to *PL2*; cf. Sec. 3.2.1). The impurity measure of each decision tree was the Gini Impurity and the depth of each decision tree was maximal, resulting in pure leaves or leaves with less than 2 samples. Feature selection based on χ^2-statistics of each attribute and comparisons of different feature sets (full set or sets with 1-20 features with best χ^2-scores) fitted to unique RFCs showed that a set of 5 features was sufficient to solve the classification task (cf. Fig. 3.1). Other studies have shown that Random Forests have good prediction quality and they can cope with a feature space including many highly correlated features (Strobl et al., 2008). Moreover, the feature importances of the RFC provide a ranking of the respective features allowing both a better linguistic interpretability of the selected

Figure 3.2: Confusion matrices showing the respective recalls of the 2 classes (*PL0* and *PL12*) in the main diagonal for the validation set (left) and the test set (right).

features (e.g., Arnold et al. (2013) and Schuppler and Schrank (2018)) and a link to the used χ^2 feature selection algorithm. In both classification experiments, a set containing 10% of each class of the training set was used for validation. Methods are evaluated by measuring the F1-score and by presenting the respective confusion matrices.

3.2.2 Results

Results of decision tree (2 prominence classes): In the binary classification task, the feature word duration in the root node of a decision tree obtained a sufficient separability, which led to a highly simplified model to distinguish between 2 prominence levels (*PL12* vs. *PL0*). If the condition word duration (w_dur) $\leqslant 0.25s$ was fulfilled, the observation was classified as *PL0*. Confusion matrices of the validation and the test set (cf. Fig. 3.2) showed that non-prominent levels had similar recalls. Prominence was recognized better in the test set (recall = 93%) and the corresponding F1-score was 92%. In both sets, non-prominent words had a F1-score > 85%.

Results of Random Forest (3 prominence classes): In the second classification task with 3 prominence levels (cf. Sec. 3.2.2), a set of 5 features (cf. Fig. 3.1) was chosen. The prominence rating of the final RFC (averaged impurity decrease of an ensemble of 1000 decision trees) showed that the feature w_dur was rated as the most important feature (cf. Fig. 3.4), followed by the features referring to the F0 range (f0_relRange and f0_Range) and the RMS range (RMS_relRange and RMS_Range). Since the relative ranges of F0 and RMS were calculated by relating the F0 or RMS range of the word to the range of the respective utterance, there was a correlation between the features f0_range and f0_relRange ($r = 0.86$ (2909), $p < .0001$) and RMS_Range and RMS_relRange ($r = 0.96$ (2899), $p < .0001$) (Kirch, 2008).

Fig. 3.3 shows the confusion matrices of the classification with three prominence levels for the test set and validation set. When comparing the recalls (corresponding to the main diagonal of the confusion matrices) of the two sets, similar results can be seen for classes *PL0* and *PL2* (recall > 82% in both cases). However, for *PL1*, a recall of only 47% was measured for the validation set, while 33% of *PL1* was predicted as *PL2*. In the test set, 19% more observations of class *PL1* were

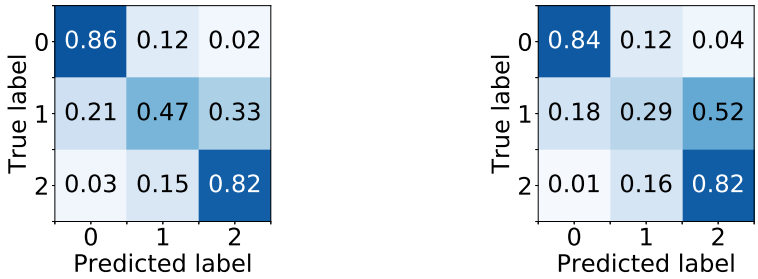


Figure 3.3: Confusion matrices showing the respective recalls of the 3 classes (*PL0*, *PL1* and *PL2*) in the main diagonal for the validation set (left) and the test set (right).

predicted as class *PL2* in comparison to the validation set. This uncertainty of classifying prominence level *PL1* was also represented in the corresponding F1-scores of the validation set (F1-score = 51%) and the test set (F1-score = 34%). No prominence was recognized in both sets with F1-score > 85%. In contrast, F1-scores of class *PL2* were higher in the validation set (F1-score = 81%) than in the test set (F1-score = 74%).

3.2.3 Discussion

Classification performance: The binary classification task indicates a good separability of prominence. A preliminary decision about word prominence could help in the annotation process by anticipating the distinction between no prominence and prominence, so the annotators can focus on a simpler manual binary decision task (between weak and strong prominence) instead of the more complex three-way decision task. Recalls of classes *PL0* and *PL2* in the second classification task with three prominence levels show very good results in both sets. Results of class *PL1* in the confusions matrices, however, indicate more variation in the production or an uncertainty in the annotation of weak prominence. One reason why the recognition performance of class *PL1* was poorer in the test than in the validation set could be that acoustic cues are weighted differently in the two data sets (cf. Sec. 3.2.1).

Contribution of acoustic features to classification and perception: In both classification methods, word duration was the most important feature for distinguishing prominence levels. Fig. 3.4 shows that prominence was represented by the classical triad of prosodic features (Lehiste, 1970) in the RFC: duration, two F0 features and two RMS features. Our experiment showed that word duration was a more important feature than F0 range, which was, in turn, more important than RMS range. These results are in line with the findings by Arnold et al. (2012). Similarly, Tamburini and Wagner (2007) found that force-accent parameters were more important for prominence in German than pitch-accent parameters. However, other studies of prominence in German found that F0 related parameters were more important (Arnold et al., 2013; Niebuhr & Winkler, 2017). Due to methodological differences, the results of the different studies are however not fully comparable (cf. Arnold et al. (2012)). Many studies on word prominence use acoustic measures that relate to the stressed vowel or syllable (e.g., Baumann and Winter (2018); Cole et al.

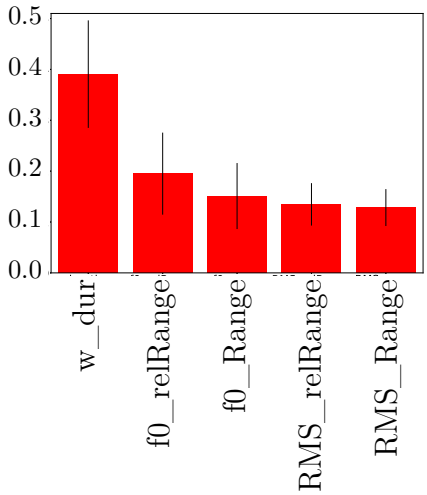


Figure 3.4: Feature Ranking of the fitted RFC corresponding to the averaged impurity decrease computed with 1000 decision trees. Black lines indicate the standard deviations referring to the impurity measurements of each tree of the forest.

(2010)), while we investigated word-level features only. Thus, we cannot conclude from the importance of word duration in our data that force-accent parameters are more important than pitch-accent parameters (cf. Tamburini and Wagner (2007)) as the stressed syllable is the constituent a force/pitch accent is associated with. Remarkably, the conclusion that duration was more important for perceived word prominence than F0 features was also drawn by one other study that investigated word prominence by measuring word-level acoustic features (Arnold et al., 2012). An advantage of our approach over studies measuring acoustic features in the stressed syllable is that it captures F0 excursions related to late peaks which are often realized outside the stressed syllable. In addition, word-level duration does not only capture whether the prominent syllable is shortened or lengthened, but also whether reductions (e.g., segment deletions) shorten non-prominent words as a whole.

The five features in the RFC all had higher values the higher the prominence level was (cf. Fig. 3.5 and Fig. 3.6). For prominence level *PL0*, word duration values are located in the lower range and are clearly distinct from those of *PL1* and *PL2*, while there is more overlap between *PL1* and *PL2* (cf. Fig. 3.5). Less scattering of *PL1* and *PL2* towards lower values in the test set (cf. Fig. 3.5) also explains why the recall of *PL12* was higher in the test set than the validation set (cf. Fig. 3.2). One reason for non-prominent words being shorter is that 94% of them were function words, which are generally shorter in terms of syllables as well as duration. Less prominent words also have a shorter duration because the speech rate in less prominent words is higher (cf. Fig. 3.5). A mixed effects model (Bates et al., 2015) with local speech rate as dependent variable, prominence rating as independent variable and word as random variable showed that the local speech rate is significantly higher for class *PL0* than class *PL1* (Est. = -3.39, t = -11.39, $p < .001$) and class *PL2* (Est. = -5.18, t = 15.78, $p < .001$). Thus, words with the same number of phones are produced faster in less prominent position.

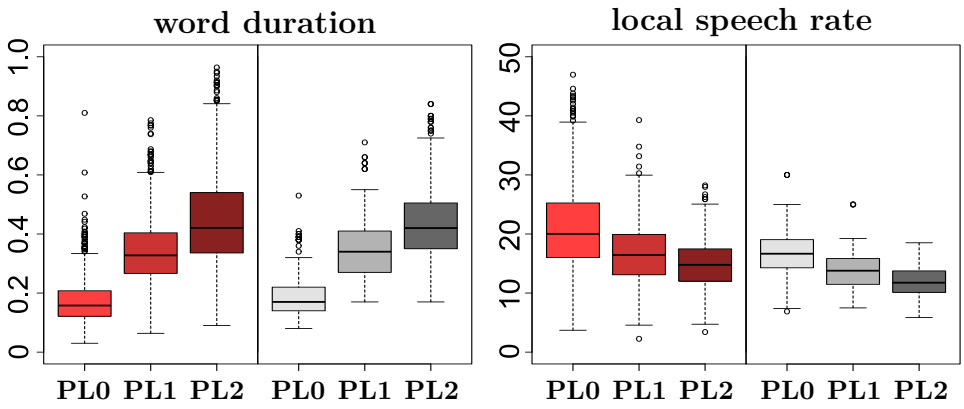


Figure 3.5: Boxplots of prominence ratings PL0, PL1, and PL2 of duration (left) and local speech rate (right), for the training set (red) and the test set (grey).

Since the range and the relative range of F0 and RMS are correlated and show similar distributions, only the relative ranges are discussed here. Fig. 3.6 shows that f0_relRange increases with prominence in both data sets. The higher F0 excursion for class *PL2* in the test set could explain the better recall of *PL12* in this set (cf. Fig. 3.2). RMS relative range also increases with prominence in both data sets. However, for RMS_relRange, the distribution of class *PL1* in the test set is similar to the distribution of class *PL2* in the training set. This difference in RMS_relRange between the two data sets could explain why level *PL1* was classified more often as *PL2* in the test set (cf. Fig. 3.3). This could be because the test set includes isolated short sentences (cf. Sec. 3.2.1), which in turn might result in a different reading behavior characterized by a different weighting of the acoustic features involved in expressing prominence.

3.2.4 Conclusions

The aim of this study was to build an annotation tool for prosodic prominence with as little pre-processing effort as possible. Therefore, the models rely on acoustic features extracted from the word and its automatically created word and phone-level segmentations, as this information is usually available first in the resource development process. We thus chose not to use any other linguistic information such as the position of the prominent syllable in the word. Based on manual prominence annotations of a small part of GRRS, we explored the combinations of different acoustic feature sets and classification methods. Our results show that both classification methods can distinguish between prominent and non-prominent words. In a binary classification, 93% of all prominent words were classified as prominent in the test set. Results with three prominence levels indicate that strong prominence is recognized well (recall = 82%), but weak prominence tends to be confused with strong prominence.

In our analysis of the contribution of acoustic features to prosodic prominence in Austrian German, word duration was the most important feature, followed by F0 range and RMS range. These results are in line with one other study of word prominence in German, but deviate from most other studies in which F0 features

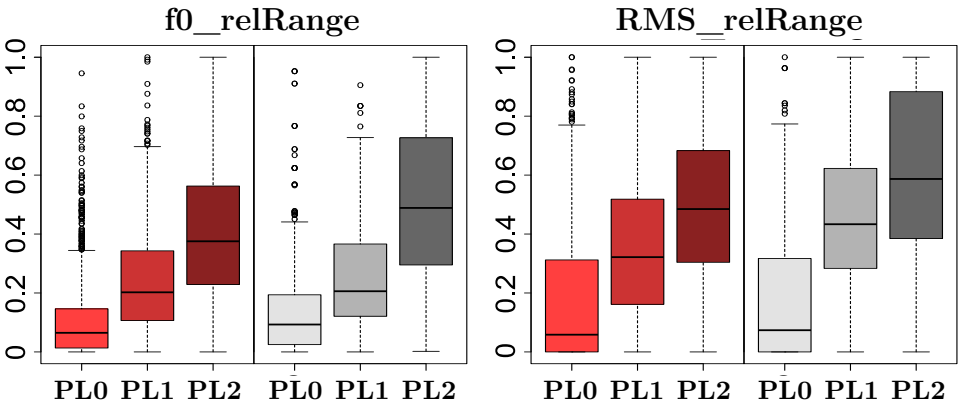


Figure 3.6: Boxplots of prominence ratings PL0, PL1, and PL2 of duration (left) and local speech rate (right), for the training set (red) and the test set (grey).

ranked higher than other acoustic features. This discrepancy could be due to different methodologies, in particular concerning the domain of acoustic feature extraction (i.e., syllable vs. word).

The presented classifiers will be used in two ways in the future: (1) as part of a semi-automatic annotation process for the rest of the read speech component of the GRASS corpus to yield faster and more consistent annotations; (2) as part of a prosody-dependent ASR system. For these two purposes, the performances reached are sufficient. Whereas the specific findings of this study will mainly be relevant for speech scientists and prosody researchers interested in German, our methodological approach of analyzing prosodic prominence from a purely acoustic perspective at the word level will also be interesting for researchers investigating the prosody of other languages.

3.3 Prominence classification in conversational speech

3.3.1 Materials and methods

GRASS corpus: This study is based on conversational speech from GRASS (GRCS) (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017). GRCS contains Austrian German conversational speech from 38 Austrian speakers, containing a total of approx. 19h of speech. Word- and phone level segmentations were created by means of a forced alignment using a Kaldi-based ASR system with a lexicon containing on average $5.57 - 6.18$ pronunciation variants per word type (Wasserfall, 2020). Phonetically trained transcribers created prosodic annotations and the resulting dataset includes a total of 5112 word tokens from 34 speakers of GRASS. The prominence annotations distinguished the prominence levels 0 (no prominence; *PL0*), 1 (weak prominence; *PL1*), 2 (strong prominence) and 3 (emphatic prominence) Prominence levels 2 and 3 were again combined (cf. Sec. 3.2.1), as emphatic prominence occurred rarely (*PL2*). Annotations were created in three stages: One annotator created a first version, which later was corrected by her/him and subsequently corrected by one of the other annotators. Based on a small subset annotated by two different annotators in those stages, the inter-annotator agreement was calculated: The overall Cohen‘s kappa was 0.72 (598 tokens), 0.72 for level 0 vs. 1 (371 tokens), 0.92 for level 0 vs. 2/3 (446 tokens) and 0.57 for level 1 vs. 2/3 (275 tokens). Other studies obtained similar agreements of 0.53 (Tamburini & Wagner, 2007) or 0.84 (Baumann & Winter, 2018).

84 basic F0 and RMS features: All features were calculated at the word-level. We calculated F0 with the library AMFM decompy (Schmitt, 2018) which includes an implementation of the pitch detection algorithm YAAPT (Zahorian & Hu, 2008). Intensity features were generated directly from the waveform by calculating the root mean square. For F0/RMS, and their respective first and second derivatives, we extracted 10 measurements: maximum, minimum, range, relative position of maximum and minimum in the word, mean, median, first and third quartile and standard deviation (60 features). Additionally, we extracted left and right slope

---

This section is based on:

[C] Julian Linke, Gernot Kubin, and Barbara Schuppler. (2023). Using word-level features for prosodic prominence detection in conversational speech. In *Proc. of ICPhS* (pp. 3101–3105).

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing). The experiments of this section are different from the original paper with respect to the following aspects: First, instead of interpolated contours we used frame-wise F0/RMS contours for calculating the entropy-based features and also applied a different normalization technique to these features. This decision was taken for a better comparison with the methods described in Sec. 4.2. Second, all cross-validation results refer to models trained on the entire feature sets. Third, the test set results differ slightly due to a different randomization process. Nonetheless, the cross-validation results demonstrate the extent to which the results can vary between different test sets. Despite these modifications, the research continues to produce findings that align closely with the original study. Finally, it is important to note that, although additional human annotations were available for the experiments, the utilized prosodic feature extractor encountered difficulties in calculating features for a substantial amount of the data. This was mainly due to characteristics in the F0 contours that could not be extracted. For this reason, a more detailed discussion of this limitation is provided in Sec. 3.3.5 (Limitations of this study).

of the maximum and minimum, absolute and relative onset and offset within the word, as well as the maximum, minimum, range and mean relative to the utterance (24 features).

12 basic durational features (DUR): We extracted word duration, phrase-level speech rate (i.e., number of segments per phrase), local speech rate (i.e., the number of segments per word duration), and relative speech rates (i.e., the ratio of the local speech rate and the minimum, maximum or median of local speech rates within a phrase). Additionally, we calculated the minimum, maximum, range, mean, median and standard deviation of local speech rates within a phrase.

4 entropy-based features: Entropy measures the spread of probability distributions and provides a measure of uncertainty of a random variable X (Cover & Thomas, 2006). If the random variable X assumes values $x_i \in \mathcal{X}$ where $\mathcal{X}$ is a finite set, the definition of entropy can be stated as

$$H(X) = -\sum_i p_i \cdot \log p_i, \tag{3.1}$$

where $p_i = Pr\{X = x_i\}$ describes the probability of X taking the value x_i, assuming that $p_i \cdot \log p_i = 0$ for $p_i = 0$.

If we observe a sequence of N (non-negative) feature values $\langle f[1], f[2], \ldots, f[N]\rangle$ within a given word, we can measure the spread of these values also by a formal entropy where the (pseudo-)probability distribution is defined by normalizing the feature values

$$p_i = \frac{f[i]}{\sum_{n=1}^{N} f[n]} \tag{3.2}$$

such that the condition for the total probability is fulfilled: $\sum_{i=1}^{N} p_i = 1$.

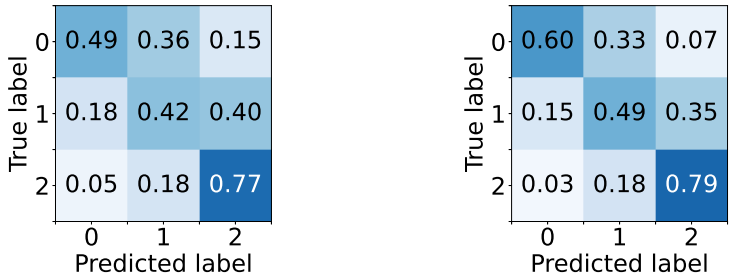
With this definition, the entropy (cf. Eq. 3.1) achieves its maximum $H_{max} = \log N$ if the feature sequence is constant $f[1] = f[2] = \cdots = f[N] = \text{const.}$, and its minimum $H_{min} = 0$ if all probabilities according to Eq. 3.2 turn out to be close to either 1 or 0, e.g., for a very non-uniform feature sequence within the given word. Note that this entropy measures the (relative) feature variability within the word, but without accounting for the time order of the feature contour. Finally, we also experimented with a normalized entropy $\tilde{H}$ obtained from division by the logarithm of the sequence length N:

$$\tilde{H} = \frac{H}{\log N}. \tag{3.3}$$

For our experiments, we applied Eq. 3.2 to the extracted F0/RMS contours[1] and calculated four additional entropy-based features (ENT) with Eq. 3.1 and Eq. 3.3 leading to two (pseudo-)entropies H (HPSF0/HPSRMS) and two log-normalized (pseudo-)entropies $\tilde{H}$ (HPSF0N/HPSRMSN) of F0/RMS.

Simulations with uniform and non-uniform distributions indicated that these entropy-based features depend primarily on the number of possible outcomes

[1]This calculation was based on frame-wise F0/RMS contours which were also extracted with the AMFM decompy library. More precisely, in this case, we used the pitch object attributes `PitchObj.samp_values` and `PitchObj.energy` to extract the F0 and RMS contours.



(a) F0+RMS (b) F0+RMS+DUR

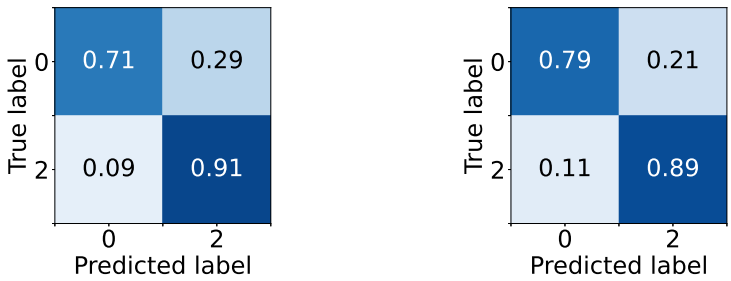
Figure 3.7: Confusion matrices (3 classes) with F0 and RMS features (a) and all features (b).

N, which in case of F0 contours corresponds to voiced segments and in case of RMS contours to word duration. Nevertheless, for words of similar lengths these measurements also encode contour variations by capturing deviations from uniform distributions (cf. a comparison of entropy-based features in Appendix A).

Random Forest: We trained Random Forest classifiers (RFCs) with the *scikit learn* toolkit (version 0.21.3) (Pedregosa et al., 2011). RFCs were built with 100 estimators, default maximum depth, a minimum samples split of 2 and the Gini impurity for measuring the quality of a split. For each of the different feature sets, we present results from two conditions, one for 2 classes (*PL0* vs. *PL2*), and one for 3 classes (*PL0* vs. *PL1* vs. *PL2*). Each classification experiment involved two steps: First, RFCs were trained with the entire feature set of a training set in order to learn about the feature's relative importance. Second, a (final) RFC was trained with the 15 most important features as given by the first step. The training and test sets were based on a random 80/20 split and we present associated F1-scores. Additionally, we provide means and standard deviations of accuracies resulting from 10-fold cross-validation experiments based on RFCs trained with the entire feature sets in order to estimate the model's generalization ability. For the latter experiments we also tested parameter robustness by comparing the original RFCs with 100 estimators to RFCs with $10, 20, 30, 40, 50$ and 500 estimators.

3.3.2 The role of durational features

In order to learn about the role of durational features for prominence classification, we conducted two classification experiments. While the first RFC was based on a selection of all 96 F0, RMS and durational features (F0+RMS+DUR) described in Sec. 3.3.1, the second RFC used a selection of the 84 F0/RMS related features only. Fig. 3.7 and Fig. 3.8 show the confusion matrices of RFCs which were trained on the basic feature set (F0+RMS+DUR) and on a subset without durational features (F0+RMS). We observe that classification performance between non-prominent and highly prominent words is high in both cases, but that non-prominent words were better classified when the RFC was trained with the basic feature set than without DUR (recall 79% vs. 71%). For *PL0* the F1-score increased form 75.2% to 77.9% by adding DUR, and for *PL2* from 89.0% to 89.2%. Corresponding cross-validation accuracies of the RFCs trained with the entire feature sets were $88\% \pm 5\%$



(a) F0+RMS (b) F0+RMS+DUR

Figure 3.8: Confusion matrices (2 classes) with F0 and RMS features (a) and all features (b)

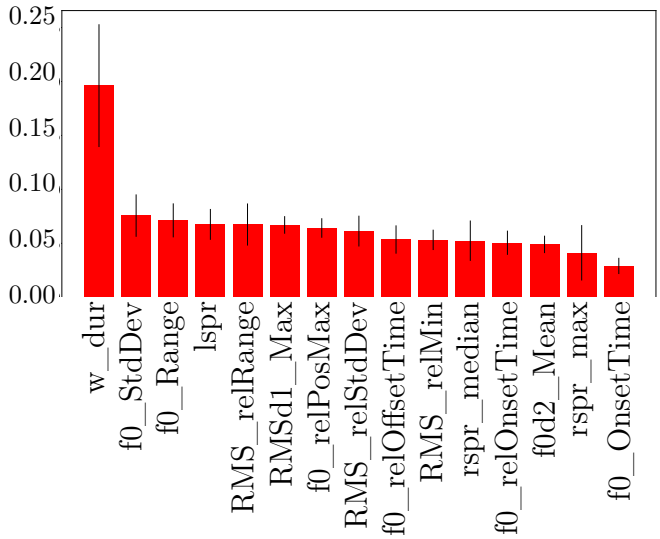


Figure 3.9: RFC feature importances for the 3 class problem with F0, RMS and DUR features.

(F0+RMS+DUR) and 85% ± 4% (F0+RMS). Furthermore, cross-validation recalls for the entire basic feature set (F0+RMS+DUR) were 81%±3% (*PL0*) and 92%±2% (*PL2*) and for the entire F0+RMS subset 72% ± 4% (*PL0*) and 92% ± 2% (*PL2*).

RFCs with 3 classes (cf. Fig. 3.7) showed a similar behaviour since the recall for *PL0* of 60.0% (F0+RMS+DUR) was higher than the recall of 49.0% (F0+RMS). However, in case of *PL1*, recalls of only 49.2% (F0+RMS+DUR) and 42.0% (F0+RMS) were achieved, while approx. 35% (F0+RMS) and approx. 40% (F0+RMS+DUR) of tokens from *PL1* were predicted as *PL2*. Respective F1-scores of *PL0*/*PL1* were 65.2%/48.8% (F0+RMS+DUR) and 54.3%/42.8% (F0+RMS). Interestingly, recalls of highly prominent words were similar in both cases (approx. 78%), indicating that *PL2* was easier to classify. In this case, cross-validation accuracies of the RFCs trained with the entire feature set were 63% ± 7% (F0+RMS+DUR) and 59% ± 4% (F0+RMS). Furthermore, cross-validation recalls for the basic feature set (F0+RMS+DUR) were 64% ± 4% (*PL0*), 47% ± 4% (*PL1*) and 77% ± 4% (*PL2*). For the subset cross-validation recalls were 57% ± 3% (*PL0*), 41% ± 3% (*PL1*) and

$76\% \pm 3\%$ (*PL2*).

Fig. 3.9 shows the feature ranking corresponding to the averaged impurity decrease of the RFC for 3 classes trained with the 15 best features (F0+RMS+DUR). Word duration (w_dur) has by far the highest importance among all features, capturing almost 20% of the overall importance. Other durational features (cf. Sec. 3.3.1), i.e. the local speech rate (lspr) or relative speech rates (rspr_median and rspr_max) were also present in the feature ranking and had similar importances as the relative relationships of the F0/RMS contours. This finding is in line with the study by Linke et al. (2020), showing that word duration is the most important feature for prominence classification in read speech (cf. Sec. 3.2.3).

Overall, durational features improve the RFC accuracy for classifying prominent words of conversational speech. Additionally, the comparison of RFCs across different estimator counts demonstrated that the model's performance was not critically dependent on the number of utilized trees, maintaining consistent cross-validation accuracies with $\geqslant 30$ estimators for the classification problems with two classes and $\geqslant 50$ estimators for the classification problems with three classes. Previous investigations on prominence cues pointed towards different trends. Whereas Cole et al. (2010) found vowel duration to be an important cue to prominence in spontaneous speech, Niebuhr and Winkler (2017) concluded that F0 and Baumann and Winter (2018) that RMS play a more important role. These studies, however, did not consider word duration, which in our experiments resulted to be the most important feature among all durational, F0/RMS features to classify prominence in conversational Austrian German.

In order to gain insights with respect to the interactions of the features, we built a cumulative link mixed model (Christensen, 2023) in R (R Core Team, 2021) with prominence as an ordinal dependent variable, the highest ranked scaled prosodic features of each category (i.e., word duration, RMS relative range and F0 standard deviation; cf. Fig. 3.9) as fixed effects and word as well as speaker identity as random effects. The model included main effects and all possible two-way interactions between fixed effects, revealing both significant main effects and interactions. Word duration emerged as the strongest predictor (Est. = 1.78, z = 29.73, $p < .001$), followed by RMS relative range (Est. = 0.31, z = 6.95, $p < .001$) and F0 standard deviation (Est. = 0.16, z = 2.55, $p < .05$). The model showed significant negative interactions between word duration and RMS relative range (Est. = -0.20, z = -4.01, $p < .001$) and between word duration and F0 standard deviation (Est. = -0.19, z = -2.45, $p < .05$), indicating that the effect of word duration diminishes for higher values of these prosodic features. Further analysis of the model indicated that non-prominent words are predicted more likely for shorter words (i.e., shorter word durations) with less prosodic variation (i.e., smaller values of RMS relative range and F0 standard deviation). This systematic pattern suggests that speakers employ a consistent strategy when marking the absence of prominence, characterized by short words with less prosodic variability. In contrast, the effect becomes less systematic for higher prominence levels, indicating that the distinction between weak and strong prominence levels relies more heavily on word duration alone. The analysis of random effects showed a substantially stronger effect for variation of word identity (standard deviation = 1.01) in comparison to variation of speaker identity (standard deviation = 0.56). Most notably, the estimated threshold coefficients describing the transitions between prominence levels showed a clear

categorical boundary between *PL0* and *PL1* (Est. = -2.72, SE = 0.13, z = -21.29), but a much weaker boundary between *PL1* and *PL2* (Est. = -0.06, SE = 0.11, z = -0.51).

3.3.3 The role of entropy-based features

To learn about the role of entropy-based features (ENT) for prominence classification, and whether they can complement phone-based durational features, we conducted two classification experiments. While the first RFC uses all 100 features (F0+RMS+DUR+ENT), the second RFC does not use any durational features (F0+RMS+ENT).

The RFC for 3 classes with F0+RMS+DUR+ENT features resulted in a large number of confusions of *PL1* (recall/F1-score: 48%/48.6%) with *PL0* or *PL2*, where approx. 14% of *PL1* was classified as *PL0* and 33% as *PL2*. In contrast, recalls/F1-scores of 60.0%/65.2% (*PL0*) and 79.1%/76.4% (*PL2*) indicated less confusions with others classes (i.e., only $3-8\%$ of non-prominent or highly-prominent words were classified as highly-prominent or non-prominent words). This result is to be expected, as also the inter-rater agreement showed to be lowest/highest for these classes. Overall cross-validation accuracies of the RFCs trained with the entire feature set reached $62\% \pm 7\%$. Furthermore, cross-validation recalls were $66\% \pm 3\%$ (*PL0*), $48\% \pm 3\%$ (*PL1*) and $75\% \pm 3\%$ (*PL2*). Compared to the RFC without ENT, the classification of weakly-prominent words improves by adding the ENT features (recall $53\% > 49\%$). The comparison with the RFC trained without any durational features (F0+RMS+ENT) indicated that developed entropy-based features compensate for durational information (similar F1-scores for classes *PL0*/*PL2* of approx. 78%/90%). With respect to the feature importances for the RFC with features F0+RMS+DUR+ENT, we observed that the four best features comprised word duration as well as the entropy features HPSF0/HPSRMS and the log-normalized entropy feature HPSF0N, which all had average importances of $> 7\%$ (capturing approx. 40% of the overall importance), while all other features had importances $< 6.4\%$.

For both the classification problems with two and three classes, prominence classification was best when adding entropy-based F0/RMS-features to the feature set. Notably, while RFCs consistently performed well across different numbers of estimators, cross-validation accuracies stabilized with fewer estimators in the classification problems with two classes ($\geqslant 10$) compared to the classification problems with three classes ($\geqslant 50$), indicating greater model robustness in the binary case. To the best of our knowledge, there exist no earlier studies on prominence classification using similar entropy-based F0 and RMS features.

3.3.4 Conclusion

This section investigated different word-level features to classify prosodic prominence, to avoid the necessity of creating manual phonetic segmentations for conversational speech. Overall, the classification performances achieved with our different sets of features were in the range of the human inter-rater agreements for the respective classes. We found that durational features (incl. speech rate variations) have a higher importance than F0/RMS features, and that among them, word duration is by far the most important feature. Experiments with entropy-based F0/RMS features

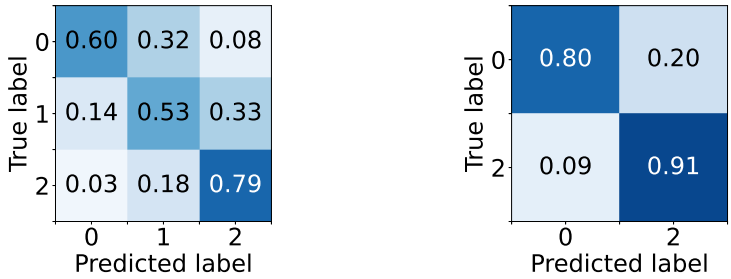


Figure 3.10: Confusion matrices from experiments with 15 best features (F0+RMS+DUR+ENT).

showed that they encode necessary durational information along with information about the features' distribution, making them useful for classifying prominence levels in conversational speech. In future, we will explore whether entropy-based F0/RMS features are also useful to capture other prosodic characteristics in speech, both with respect to speech analysis as well as in ASR.

3.3.5 Limitations of this study

The prosodic feature extractor encountered difficulties in calculating features for a substantial portion of the data, despite the availability of additional human annotations. The main issues were:

1. Differences between human-annotated prosodic word boundaries and forced-aligned word boundaries which allowed for the alignment of only 70% of the data (cf. Sec. 4.5.2).

2. For the remaining *word-aligned* data, the peak detector failed to extract all characteristics in the F0 contours for approx. half of the data. More precisely, the pitch detector failed to find only F0 valleys (minimum) in approx. 10.5% of cases, only F0 peaks (maximum) in approx. 18% of cases, and both F0 peaks/valleys in approx. 21.5% of cases which made it impossible to calculate certain prosodic features accurately (i.a., relative position of F0 maximum and minimum in the word).

Further analysis revealed that the pitch extractor often failed when word duration was small (approx. 0.15 s) or when the standard deviation of F0 was small (approx. 2.5 Hz). For word duration of approx. > 0.2 s or F0 standard deviation of approx. > 5 Hz, it was generally feasible to calculate prosodic features for the F0 contours. These findings suggest difficulties in pitch extraction and pitch detection especially for shorter utterances or for utterances of less F0 variation. These limitations are particularly relevant for conversational speech, where a large portion of the data consists of short utterances (cf. Fig. 4.4a), stemming from the lively turn-taking. While these characteristics make prosodic feature extraction challenging, they are inherent to the nature of spontaneous dialogue, highlighting the need for robust methods in analyzing such data, as for instance the use of entropy-based features presented here which do not rely on algorithms to detect maxima and minima in F0.

Therefore, we explored alternative approaches, such as chroma features, which do not depend on F0 extraction or phone segmentations and encompass aspects of all DUR+F0+RMS features in a lower-dimensional representation (Linke et al., 2025). These findings demonstrated, that chroma features achieved comparable performance to classical prosodic features, without the drawback of data loss encountered with traditional methods.

Chapter 4

Automatic speech recognition

4.1 Introduction

In recent years, we have observed a rapid advancement of Automatic Speech Recognition (ASR) architectures resulting in increasingly improved performance across various benchmarks (cf. Gabler et al. (2023) for an overview). Especially for more spontaneous speaking styles like conversational speech (CS) there has been an increasing interest to improve performance for two main reasons: 1) Human-machine interaction with social robots or speech agents is becoming an integral part of our everyday lives (e.g., virtual assistants like Amazon's Alexa, Apple's Siri, Google Assistant, Microsoft's Cortana or speech recognition applications for chatbots like ChatGPT). 2) The need for applications able to generate high-quality automatic transcriptions for spontaneous conversations between two or more humans (e.g., transcriptions of meeting recordings) in various domains, which may be particularly useful for humans with hearing, speaking or visual disabilites. Hence, the rapid advancements in speech technology should address the increasing demand for machines to better adapt to human interactional communication behaviour (European Union, 2024). This demand also highlights the necessity for the development of highly performing ASR systems for, from a modelling point of view, complex conversational speech that occurs in our, from a human point of view, "simplest" every-day communications.

Despite this necessity for ASR to perform well on conversational speech, most benchmarks databases mainly contain read, prepared or well pronounced speech (e.g., Librispech (Panayotov et al., 2015), Common Voice (Ardila et al., 2020), Multilingual Librispeech (Pratap et al., 2020)). Likewise, there seems to be an overly optimistic bias towards interpreting ASR performances through existing benchmarks when comparing published Word Error Rates (WERs) (Szymański et al., 2020). As a result, the best WERs for the frequently used Switchboard corpus (J. J. Godfrey et al., 1992), a corpus of the well-resourced language American

---

This introduction has been reformatted from:

[D] Julian Linke, Bernhard C. Geiger, Gernot Kubin, and Barbara Schuppler. (2025). What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures. *Computer Speech and Language, Volume 90, 101738.*

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

English, are in the range of 4.3% (Tüske et al., 2021) to 5.1% (Xiong et al., 2018). In contrast, the OpenASR21 challenge demonstrates that WERs for low-resource languages fall within the range of 32% (Swahili) to 68% (Farsi) (Peterson et al., 2022). At the same time, it is not straightforward to sufficiently define sub-categories of spontaneous speaking styles, which complicates direct comparisons of reported WERs for different speaking styles (cf. Linke, Wepner, et al. (2023) for a detailed description and categorization of speaking styles). As a consequence, conversational speech is often subsumed in, or reduced to, spontaneous speech, which may also lead to an overly optimistic picture. For example, the Switchboard corpus certainly contains spontaneous speech, but since the recorded conversations took place over the telephone, several characteristics of conversational speech that complicate automatic speech recognition are not represented in the corpus.

There is a general requirement for ASR systems to be robust for different speaking styles and/or different speakers. Modern ASR architectures based on self-supervised pre-training provide powerful solutions especially in low-resource scenarios (Baevski, Zhou, et al., 2020; Conneau et al., 2021), and the broad study on domain shifts in self-supervised pre-training by (Hsu et al., 2021) demonstrated how pre-training on more domains improves robustness in general. Our previous studies on low-resource speech recognition with conversational Austrian German reinforced the general effectiveness of fine-tuning a pre-trained cross-lingual speech representation model; the results, however, indicated a lack of robustness with respect to speaker-dependent and conversation-dependent WER distributions (Linke et al., 2022; Schuppler, Hagmüller, et al., 2014). Surprisingly, we also also discovered that this lack of robustness is not affected by the amount of utilized training data (Linke et al., 2022). This is evident as the reported conversation-dependent WERs in low-resource scenarios had mean values and standard deviations of $56.19 \pm 5.4\%$ and $57.28 \pm 6.26\%$. In a fine-tuning scenario, these WERs were reduced to $25.06 \pm 4.42\%$, i.e., the standard deviation computed over different conversations remained at a similar level as without fine-tuning.

4.1.1 What makes ASR on conversational speech so complex?

In contrast to read speech, conversational speech is inherently more complex because its production has an entirely different nature which originates from the fact that it is planned and produced by conversational interlocutors together in real-time (Lopez et al., 2022). Characteristics of conversational speech are that interlocutors frequently produce grammatically incomplete or even grammatically wrong utterances, self-interruptions, backchannels and that they exhibit disfluencies and speak with a high degree of acoustic reduction and pronunciation variation (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014). As a result, these spontaneous interactions show complex inter-speaker and intra-speaker variation which, among other things, is also reflected in the speaker's attitude towards the listener (Wright, 2006).

In the broader field of speech recognition, the effects on WERs have been studied from different perspectives. In general, (Lopez et al., 2022) found for conversational speech from three different languages that ASR systems struggle with basic communicative events such as conversational non-lexical tokens (e.g., backchannels or delay markers) and word segments which emerged from self-correction. An older study from (Hirschberg et al., 2004) revealed that the prosodic features relating to

F0 excursion, loudness, and longer duration are significant predictors for recognized and misrecognized utterances of (task-oriented) spoken dialogue systems. A study on English telephone conversations found that words with low intensity, high F0 value or shorter duration tend to be more often misrecognized than words of high intensity, low average word level F0 and longer duration (Goldwater et al., 2008). (Goldwater et al., 2008) also mentioned that the large individual differences across speakers with respect to WERs might be an indicator for why their ASR systems did not adapt well to prosodic variation within and across speakers. More recently, but still with respect to an HMM-based ASR system, the study by (Wepner et al., 2022) analyzed the effect of prosodic characteristics (i.e., phrase boundaries, position in the phrase, prominence level and stress accent type) on WERs for conversational Austrian German and confirmed that words with longer durations were recognized correctly more often then shorter words, and that WERs were significantly lower for prominent words.

We emphasize that the challenges posed by conversational speech on ASR do not only originate in the characteristics that are intrinsically related to the variation in the speech itself, but that moreover the manual transcription quality, the reference labels, contain more errors and exhibit lower agreement across different annotators. In comparison to read speech, where a a reference text is the basis for the produced speech signal, in spontaneous speech the reference text was transcribed from the given signal. (Gabler et al., 2023) hypothesized that this annotation "problem" can be viewed as a causality problem and should be considered in ASR architectures. Hence, comparing and analyzing WERs of spontaneous speaking styles should always be viewed with caution since human word errors cannot be ruled out even for professional transcribers.

4.1.2 GMM-HMM/DNN-HMM versus transformer-based ASR

ASR as a research field has a long history of approx. 70 years beginning with a first publication on a single-speaker digit recognition task (Davis et al., 1952). The field has seen a multitude of approaches and methodologies, each with its unique strengths and challenges. Since our work aims at investigating how different ASR architectures deal with different characteristics of conversational speech, we provide an overview on the three most influential approaches: Hidden Markov Models with Gaussian mixture models (GMM-HMM), Hidden Markov Models with Deep Neural Nets (DNN-HMM) and ASR architectures based on Transformers (i.e., with self-supervised learning or sequence-to-sequence learning).

In general, ASR architectures should predict the optimal word sequence $\mathbf{W}$ given a spoken speech signal $\mathbf{X}$. In this case, the optimality condition relates to maximizing the *a posteriori* probability (MAP) of the word sequence:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P(\mathbf{W}|\mathbf{X}). \tag{4.1}$$

The posterior probability $P(\mathbf{W}|\mathbf{X})$ can either be modelled directly or be transformed with Bayes' Rule by formulating the equivalent problem

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{X}|\mathbf{W})P(\mathbf{W}), \tag{4.2}$$

where the likelihood $P(\mathbf{X}|\mathbf{W})$ is calculated by an *acoustic model* (AM) and the prior $P(\mathbf{W})$ is calculated by a *language model* (LM). As an example, the prior probability $P(\mathbf{W})$ of a specific word sequence can be estimated with simple n-grams (Shannon, 1948) but more recent LMs are typically based on deep neural nets (e.g., LSTMs (Hochreiter & Schmidhuber, 1997) or transformers (Vaswani et al., 2017)). In the subsequent paragraphs we focus on providing an overview of significant acoustic modelling approaches.

There is a long tradition of ASR architectures based on GMM-HMM which are implemented in tools like HTK (Young et al., 2002) or Kaldi (Povey et al., 2011). In principle, a formulation for the AM likelihood is

$$\begin{aligned} P(\mathbf{X}|\mathbf{W}) &= \sum_{\mathbf{S}} p(\mathbf{X}|\mathbf{S}) P(\mathbf{S}|\mathbf{W}) \\ &= \sum_{\mathbf{S}} \left[p(\mathbf{X}|\mathbf{S}) \prod_{l=1}^{L} P(\mathbf{s}^{(w_l)}|w_l) \right] \\ &= \sum_{\mathbf{S}} \left[\sum_{\boldsymbol{\theta}} \left(p(\boldsymbol{\theta}, \mathbf{X}|\mathbf{S}) \right) \prod_{l=1}^{L} P(\mathbf{s}^{(w_l)}|w_l) \right] \\ &= \sum_{\mathbf{S}} \left[\sum_{\boldsymbol{\theta}} \left(P(\theta_1|\theta_0) \prod_{t=1}^{T} p(\mathbf{x}_t|\theta_t) P(\theta_{t+1}|\theta_t) \right) \prod_{l=1}^{L} P(\mathbf{s}^{(w_l)}|w_l) \right], \end{aligned} \tag{4.3}$$

where $\mathbf{S}$ is a particular sequence of pronunciations with each $\mathbf{s}^{(w_l)}$ being a valid pronunciation sequence for word w_l, and where $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots \theta_T$ is a state sequence with Markovian dynamics (hence HMM) through the composite model, with θ_0 and θ_{T+1} being non-emitting entry and exit states (Gales & Young, 2007). The continuous output density is modeled with a GMM where the state output density of observation $\mathbf{x}_t$ for a specific state j is

$$p(\mathbf{x}_t|\theta_t = j) = \sum_i c_i^{(j)} \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_i^{(j)}, \boldsymbol{\Sigma}_i^{(j)}), \tag{4.4}$$

with the weight $c_i^{(j)}$ for Gaussian component i, mean $\boldsymbol{\mu}_i^{(j)}$ and covariance matrix $\boldsymbol{\Sigma}_i^{(j)}$. The estimation of the AM parameters for transitions $P(\theta_{t+1}|\theta_t)$ and emissions $p(\mathbf{x}_t|\theta_t)$ can be solved with the forward-backward algorithm (Rabiner, 1989) and the best state sequence can be estimated by applying the Viterbi algorithm (Forney, 1973). Modeling probability distributions with GMMs has several advantages. For instance, with enough components, GMMs can model probability distributions to any required level of accuracy and they can be conveniently fitted to the data using the expectation-maximization algorithm (Hinton et al., 2012). In general, the training of ASR systems with GMM-HMM are based on several training stages: The first training stage involves a *monophone system* while subsequent stages are typically based on *triphone systems* which can be additionally improved by applying several feature transformations (e.g., LDA, MLLT (Gopinath, 1998), fMLLR (Gales, 1998), etc.).

With the advent of deep learning, context-dependent DNNs replace the GMMs for acoustic modeling. Hence, instead of modeling the likelihood $p(\mathbf{x}|\theta)$ and computing the MAP state sequence using the Viterbi algorithm, DNNs directly model the

posterior $p(\theta|\mathbf{x})$ to allow for a discriminative estimation of θ. Note that the DNN can take a context window around an input vector $\mathbf{x}$. For example, in order to estimate the posterior distribution of realizations (e.g., monophones, triphones or senones) for θ_t, the DNN might consider five frames: two preceding ($\mathbf{x}_{t-2}$ and $\mathbf{x}_{t-1}$), the current frame $\mathbf{x}_t$, and two subsequent ones ($\mathbf{x}_{t+1}$ and $\mathbf{x}_{t+2}$). Furthermore, the shift from a generative framework (GMM-HMM) to a discriminative framework (DNN-HMM) enables the exploitation of deep architectures including multiple non-linear transformations. Hence, this framework enables hierarchical feature learning which has yielded substantial improvements in ASR performance.

In order to train a DNN-HMM system, a GMM-HMM system first generates forced alignments (FA). This means that the GMM-HMM model structure is also reused to specify the HMM topology and the phone set in the output layer. One possible training criterion is the cross-entropy between the reference posterior distribution $P_{\text{ref}}(\theta = s|\mathbf{x}_t)$ which is given by the FAs and the predicted posterior distribution $P(\theta = s|\mathbf{x}_t)$, i.e., one aims to minimize

$$\mathcal{F}_{\text{CE}} = -\sum_{t=1}^{T}\sum_{s=1}^{N} P_{\text{ref}}(\theta = s|\mathbf{x}_t)\log\left[P(\theta = s|\mathbf{x}_t)\right], \tag{4.5}$$

where N describes the number of possible realizations (usually tens of thousands). This function can be simplified in the standard cross-entropy case where the reference posterior distribution is represented as a one-hot vector leading to the negative log likelihood

$$\mathcal{F}_{\text{CE}} = -\sum_{t=1}^{T}\log\left[P(\theta = s_t|\mathbf{x}_t)\right], \tag{4.6}$$

where s_t describes the reference senone obtained by FA at time t.

Since speech recognition is a sequence classification problem, the frame-based objective function $\mathcal{F}_{\text{CE}}$ is not optimal. For that reason, there exists another well-known solution for training hybrid DNN-HMM systems which is a lattice-free version of the maximum mutual information (MMI) criterion, also known as LF-MMI (Povey et al., 2016). In general, the maximum mutual information objective is given as

$$\mathcal{F}_{\text{MMI}} = \sum_{u=1}^{U}\log\left[\frac{p(\mathbf{X}_u, \mathbf{W}_u)}{p(\mathbf{X}_u)P(\mathbf{W}_u)}\right] = \sum_{u=1}^{U}\log\left[\frac{p_\lambda(\mathbf{X}_u|\mathbf{W}_u)^\kappa P(\mathbf{W}_u)^\kappa}{\sum_{\mathbf{W}} p_\lambda(\mathbf{X}_u|\mathbf{W})P(\mathbf{W})^\kappa}\right], \tag{4.7}$$

where U denotes the number of training utterances, λ the parameters of the AM (i.e., parameters for GMM-HMM systems or DNN-HMM systems) and κ a probability scale (Bahl et al., 1988; Povey, 2003). However, in case of LF-MMI for DNN-HMM systems the numerator and denominator of Eq. (4.7) need to be approximated by the *forward-backward* algorithm leading to the objective

$$\mathcal{F}_{\text{LF-MMI}} \approx \sum_{u=1}^{U}\log\left[\frac{P(\mathbf{X}_u|\mathbb{G}_{num})}{P(\mathbf{X}_u|\mathbb{G}_{den})}\right], \tag{4.8}$$

where $\mathbb{G}_{num}$ denotes a composite numerator HMM graph including all possible state sequences for one training transcription $\mathbf{W}_u$ (numerator graph) and $\mathbb{G}_{den}$ a denominator HMM graph including all possible state sequences of all possible word

sequences (denominator graph) (Hadian et al., 2018; Tian et al., 2023). In general, the LF-MMI objective tries to make the correct transcription of the numerator graph $\mathbb{G}_{num}$ more probable relative to a large space of possible transcriptions given by the denominator graph $\mathbb{G}_{den}$. Note that the LF-MMI objective function also relies on an explicit segmentation of the input sequence, created for instance by means of a forced alignment.

Most recent advances in ASR have been achieved by means of attention-based models that incorporate a time-dependent scoring mechanism in order to better capture complex sequential dependencies between the audio input and text output. First, attention-based models were introduced in the field of machine translation by including an alignment model into a *RNN Encoder-Decoder* framework (Bahdanau et al., 2015). Some years later (Vaswani et al., 2017) proposed the *transformer*, which in its original form is again an encoder-decoder framework, but now relies entirely on *attention mechanisms* to optimally capture global dependencies between the input and the output sequences.

In the field of speech recognition Baevski, Zhou, et al. (2020) proposed wav2vec2 which is based only on the encoder part of the Transformer and thus enables training of contextualized speech representations. This recent model is of particular interest in the field, because it learns powerful speech representations from raw audio data (pre-training) followed by fine-tuning on transcribed speech data. Surprisingly, this system demonstrates outstanding performance on the Librispeech (Panayotov et al., 2015) test set when fine-tuned with just ten minutes of in-domain transcribed speech data with WERs of 4.8%/8.2% (clean/other). One reason for this outstanding performance is attributed to the objective function utilized during self-supervised pre-training. This function is described by a composite loss

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d, \tag{4.9}$$

where $\mathcal{L}_m$ represents a contrastive loss and $\mathcal{L}_d$ denotes a diversity loss. Within this framework, the contrastive loss is of particular interest since it aims to identify the true quantized latent speech representation $\mathbf{q}_t$ given the contextualized representation $\mathbf{c}_t$ in a pool of quantized candidate representations $\tilde{\mathbf{q}} \in \mathbf{Q}_t$. This pool includes the discrete representation $\mathbf{q}_t$ along with several distractors. Simultaneously approx. 49% of all time steps are masked. Hence, wav2vec2 effectively uses the InfoNCE loss (van den Oord et al., 2018) in order to maximize the similarity between a contextualized representation $\mathbf{c}_t$ and a discretized representation $\mathbf{q}_t$ which has the practical advantage that negative samples do not need to be sampled from the same category as the positive samples (Mohamed et al., 2022). As a result, the contrastive loss can be formulated as

$$\mathcal{L}_m = -\log \left[\frac{\exp\left(S_c(\mathbf{c}_t, \mathbf{q}_t)\right)}{\sum_{\tilde{\mathbf{q}} \in \mathbf{Q}_t} \exp\left(S_c(\mathbf{c}_t, \tilde{\mathbf{q}})\right)} \right], \tag{4.10}$$

where $S_c(\cdot)$ denotes the cosine similarity between two vector representations. After pre-training, the initialized wav2vec2 model can be adapted for several downstream tasks. For speech recognition, the learned representations can be fine-tuned using labeled data by minimizing the CTC loss (Graves, Fernandez, et al., 2006). This

loss is given by the objective function

$$\mathcal{F}_{\text{CTC}} = -\log\left[\sum_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}|\mathbf{X})\right] = -\log\left[\sum_{\mathbf{s}\in\mathbf{S}}\prod_{t=1}^{T} P(\theta = s_t|\mathbf{x}_t)\right], \tag{4.11}$$

where $\mathbf{S}$ denotes the set of all valid alignment sequences $\mathbf{s}$ (including blank tokens and repetitions). The CTC objective function depends only on the sequence of labels and not on their segmentation. Hence, there is no explicit segmentation of the input sequence necessary. Another recent ASR architecture which is built upon the entire Encoder-Decoder Transformer framework is Whisper (Radford et al., 2023). In contrast to wav2vec2, which relies on pre-training in order to enable fine-tuning for a downstream task like speech recognition, Whisper was trained under weak supervision on 680 000 h of multilingual audio from the internet. Because of the Encoder-Decoder architecture of Whisper, the system simultaneously learns powerful contextualized representations of speech through the encoder together with auto-regressive mappings for sequence generation via the decoder. The transcripts of Whisper's training speech data only included the raw text, and all audio files were segmented into 30 s chunks. Rather than extracting features from the raw audio (e.g., in comparison to wav2vec2), Whisper relies on log-magnitude Mel Spectrogram features with a frame length of 25 ms and a frame shift of 10 ms. These features are subsequently processed through two convolutional layers with GELU activation functions (Hendrycks & Gimpel, 2016) before being forwarded to the Transformer network. The zero-shot results of Whisper on English indicate a close to human-level performance when compared to transcriptions given by professional human transcribers and large improvements in robustness when compared to a supervised model trained only on Librispeech (Radford et al., 2023).

4.2 What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures

4.2.1 Motivation

This work aims to analyze WER distributions in detail for different HMM- and transformer-based ASR architectures in order to better comprehend the robustness problem, which appears particularly problematic in the case of conversational speech recognition and seems to be linked to substantial differences in prosody, pronunciation, and utterance lengths. While ASR architectures continuously improve over time largely due to the success of memory-based systems (e.g., based on transformer architectures (Vaswani et al., 2017)) trained on massive amounts of data, a fundamental question remains: Does the challenge of automatically recognizing conversational speech naturally resolve itself with time and more data? Given the high degree of variation in WERs, it is not to be expected that such challenges disappear on their own accord. Given our earlier mentioned experiments and experience with conversational speech, we do not believe that this problem will resolve itself by only using more data. Analyzing conversational speech recognition results of different ASR architectures does not only seem a valuable, but even a mandatory step towards a better understanding of the ASR challenges resulting from this speaking style.

While previous studies have investigated the effects of various factors on WERs in ASR (i.e., prosodic features or communicative events), they have largely overlooked the impact of dialectal pronunciation variation. In contrast to formal, task-oriented dialogues, casual speech corpora are more likely to exhibit pronunciation variation due to dialectal differences which may significantly affect ASR performance (Linke, Wepner, et al., 2023). This distinction highlights the importance of considering the interplay between speaking style and variety when evaluating ASR systems, particularly in informal conversational settings (Linke, Kadar, et al., 2023).

This study extends the existing literature on WER analyses of ASR systems in the following directions:

- We perform analyses for several ASR systems, i.e., hybrid DNN-HMM models and transformer-based models.
- We consider how these architectures were leveraged for training with respect to the utilized speech data (e.g., with in-domain data or without in-domain data or with both in-domain and out-of-domain data).
- We take into account the learning strategy (e.g., learning based on training stages and modules or end-to-end training).

This section has been reformatted from:

[D] Julian Linke, Bernhard C. Geiger, Gernot Kubin, and Barbara Schuppler. (2025). What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures. *Computer Speech and Language, Volume 90, 101738.*

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

Table 4.1: Overview of all possible combinations of the three aspects **A1** (HMM (H) vs. transformer-based (T)), and **A2** (trained or fine-tuned on data from the target language and style) and **A3** (incorporation of explicit linguistic knowledge). Unfeasible combinations (i.e., HMM-based systems without explicit linguistic knowledge) and combinations which were not part of this study are grayed out (e.g., Kaldi or wav2vec2/Whisper trained or fine-tuned on out-domain and data with explicit linguistic knowledge).

A1	**A2**	**A3**	**ASR System Description**
H	–	–	Unfeasible
H	–	✓	e.g., Kaldi trained on out-domain data and with explicit linguistic knowledge
H	✓	–	Unfeasible
H	✓	✓	Kaldi trained on in-domain data and with explicit linguistic knowledge
T	–	–	Whisper fine-tuned/trained on out-domain data and without explicit linguistic knowledge
T	–	✓	e.g., wav2vec2/Whisper fine-tuned on out-domain data and with explicit linguistic knowledge
T	✓	–	wav2vec2 fine-tuned on in-domain data and without explicit linguistic knowledge
T	✓	✓	wav2vec2 fine-tuned on in-domain data and with explicit linguistic knowledge

- Whereas most studies considering the complexity of conversational speech analyze ASR performance exclusively on the word level, we consider the utterance level in order to capture important information coming from the sequential characteristics of conversational speech (as inspired by Hirschberg et al. (2004)).
- The design of our WER analysis allows to untangle which factors on WER stem from conversational speech characteristics, and thus transfer to other languages, and which stem from variation related to the distance of the regional variety to the standard pronunciation.

We emphasize that the challenges posed by conversational speech on ASR do not only originate in the characteristics that are intrinsically related to the variation in the speech itself, but that moreover the manual transcription quality contain more errors and exhibit lower agreement across different annotators. In comparison to read speech, where a reference text is the basis for the produced speech signal, in spontaneous speech the reference text was transcribed from the given signal. Gabler et al. (2023) hypothesized that this annotation "problem" can be viewed as a causality problem and should be considered in ASR architectures. Hence, comparing and analyzing WERs of spontaneous speaking styles should always be viewed with caution, since human word errors cannot be ruled out even for professional transcribers.

4.2.2 Design of this study

The main aim of this study is to gain insights about which aspects of casual, conversational speech cause the largest challenges for different ASR architectures. Specifically, we are interested in analyzing the effects of the following aspects:

A1) HMM vs. transformer-based,

A2) amount of training data from the target language and style, and

A3) incorporation of explicit linguistic knowledge.

For an overview, Tab. 4.1 describes all possible combinations of these three aspects ($2^3 = 8$ aspect combinations). Unfeasible combinations (i.e., HMM-based systems

without explicit linguistic knowledge) and combinations which were not part of this study are grayed out (e.g., Kaldi or wav2vec2/Whisper trained or fine-tuned on out-domain data and with explicit linguistic knowledge). For this purpose, we chose four ASR systems that are distinct with respect to three aspects: The first ASR system is Whisper, which provides a zero-shot multilingual ASR system (Radford et al., 2023) trained on 680 000 h of multilingual and multitask speech data collected from the web. We use this system as a representative for ASR architectures that are 1) transformer-based **(Relates to A1)**, 2) do not require any data from the target language/style **(Relates to A2)** and 3) do not have any explicit linguistic knowledge **(Relates to A3)**. The second ASR system is based on the Kaldi framework (Povey et al., 2011), which we train on 19h of conversational Austrian German (i.e., the GRASS corpus, Schuppler, Hagmüller, et al., 2014). We use this system as a representative for ASR architectures that are 1) HMM-based **(Relates to A1)**, 2) are trained with merely a small amount of data from the target language and style **(Relates to A2)**, and 3) have explicit linguistic knowledge incorporated in the form of the AM, a pronunciation lexicon with multiple variants per word and an n-gram LM **(Relates to A3)**. The third ASR system is based on the wav2vec2 framework (Baevski, Zhou, et al., 2020). Its XLSR model is pre-trained on 56 000 h of multilingual speech data (Conneau et al., 2021), which we fine-tune with the above mentioned GRASS corpus. We use this ASR system as representative for ASR architectures that are 1) transformer-based **(Relates to A1)**, 2) fine-tuned on small amounts of data from the target language and style **(Relates to A2)**, but 3) do not have any explicit linguistic knowledge **(Relates to A3)**. Since the wav2vec2 architecture enables also a decoding strategy including a lexicon and LM, we use that mode of wav2vec2 as our fourth ASR system, representing ASR systems that are 1) transformer-based **(Relates to A1)**, 2) fine-tuned on target domain data **(Relates to A2)** and 3) have explicit linguistic knowledge **(Relates to A3)**.

We chose a zero-shot version of Whisper as a comparative benchmark leading to one system without explicit linguistic knowledge. This choice was motivated by the fact that the zero-shot Whisper model achieves near-human-level accuracy in English but simultaneously performs even better in German with respect to Multilingual Librispeech ($5.5\% < 6.2\%$) or Common Voice 9 ($6.4\% < 9.5\%$) (Radford et al., 2023). Thus, instead of a comparison with an instance of Whisper fine-tuned on the target data, we included fine-tuned models based on the wav2vec2 architecture which also enables different decoding strategies. More precisely, we decided to compare w2v (with only implicit linguistic knowledge in the AM) and w2vLM (with implicit linguistic and explicit linguistic knowledge). Additionally, as a low-resourced representative of HMM-based architectures, we included a Kaldi system, which compared to the transformer-based architectures, allows the integration of a pronunciation lexicon with multiple entries per word type. This choice was also motivated by the evaluation in Linke, Wepner, et al. (2023) which reveals that the acoustic and language modeling of a low-resourced Kaldi system for Austrian German benefits only slightly from additional speech data from other spontaneous German corpora with absolute mean WER improvements of approx. $1\% - 2\%$. For the four remaining ASR systems, this study contributes:

1. A comparison of conversation-dependent WERs of all architectures while focusing on the general robustness problem and speaker-pair dependency.

2. The overall correlation of conversation-dependent WERs of all architectures with acoustic, pronunciation and perplexity features that capture important characteristics of conversational speech.

3. A more detailed statistical analysis of WERs by means of *Interaction Forests* (Hornung & Boulesteix, 2022a) to show how the different ASR architectures are affected by specific characteristics of the utterances (i.e., combinations of the features).

The structure of this study is as follows: Sec. 4.2.3 describes all materials used for this study. Sec. 4.2.4 presents the experimental results for the four different ASR systems. Sec. 4.2.5 presents the features that represent variation in casual, conversational speech and that show high correlations with the conversation-dependent WERs. Sec. 4.2.6 then provides a detailed statistical analysis of how the different characteristics of conversational speech affect conversation-dependent WERs in the different ASR architectures by means of the recently introduced *Interaction Forests* (Hornung & Boulesteix, 2022a).

4.2.3 Materials

All experiments in this study are based on the GRASS corpus (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017), which contains a total of approx. 30 h of Austrian German read speech (RS) and conversational speech (CS) from 38 speakers (19 female and 19 male). All GRASS speakers were born in the same broad dialectal region (Eastern Austria), had been living in an urban area for several years and had a higher education degree. Despite controlling for these extralinguistic factors, studies have provided evidence that there is a high degree of pronunciation variation in GRASS (Geiger & Schuppler, 2023).

For ASR experiments with RS, we utilized the RS component and the command component from the GRASS corpus. In case of the RS component, all speakers read phonetically balanced sentences, whereas in the command component, the same speakers read commands and keywords. For both components, we normalized all numbers to text. Overall, for the RS experiments, we used 4322 utterances (approx. 117 utterances per speaker, 3767 utterances for the RS component and 565 utterances for the commands component) and a total of approx. 4.7 h of speech data after pre-processing. The mean utterance duration and standard deviation of the RS speech data was $3.9\,\text{s} \pm 1.5\,\text{s}$ and the mean number of tokens per utterance and standard deviation was 5.64 ± 3.85 (cf. Tab. 4.2).

One important characteristic of GRASS is that the speakers of the RS component and command component are the same as in the CS component, and that RS and CS were recorded in the same studio with the same equipment. It is common knowledge in the ASR community that speaker identity and recording quality are among the largest factors that affect ASR performance. Using GRASS, we can be sure that observed performance differences between RS and CS are related only to differences in speaking style.

The CS component contains conversations between pairs of persons who have known each other for several years and were either friends, couples or family members. In general, each speaker pair was recorded for one hour without interruption which allowed a fluent and highly spontaneous conversation. During the recording of

Table 4.2: Overview of the used Austrian German speech data (GRASS) after pre-processing, separately for read speech (RS) and conversational speech (CS). In this table, **utts** stands for utterances, **tkns** for tokens, **spks** for speakers, and **convs** for conversations.

GRASS	hours	#utts	utterance durations	#tkns	#spks	#convs
RS	4.7 h	4322	$3.9\,\text{s} \pm 1.5\,\text{s}$	5.64 ± 3.85	37	-
CS	14.4 h	33734	$1.54\,\text{s} \pm 1.42\,\text{s}$	5.78 ± 6.11	38	19

the CS component, there was no experimenter present and there was no restriction with respect to dialogue topics or speaking behaviour which led to natural and partially dialectal pronunciation including characteristics typical for conversational speech (e.g., laughter, the use of swear words or regularly occurring overlapping speech (Schuppler et al., 2017)). For all experiments, we excluded utterances containing laughter, singing, imitations/onomatopoeia, unintelligible word tokens (as tagged by the annotators) and artefacts (e.g., accidental touch of the microphone), leaving us with approx. 14.4 h of CS data for the ASR experiments. Additionally, we standardized the reference text to lowercase, removed punctuations and unified different backchannel labels (`hm, hmm, mh, mhh, mmh, mhm`) to `mhm`. Overall, the resulting CS data contains 33734 utterances (approx. 1776 audio files per conversation). The corresponding mean utterance duration and standard deviation of the CS speech data was $1.54\,\text{s} \pm 1.42\,\text{s}$ and the mean number of tokens per utterance and standard deviation was 5.78 ± 6.11 (cf. Tab. 4.2).

4.2.4 ASR Experiments

This section presents the ASR experiments on GRASS with respect to the three different ASR architectures Whisper, Kaldi and wav2vec2 with and without a lexicon/LM (w2v and w2vLM). We provide speech recognition results for both the GRASS RS (GRRS) component and the GRASS CS (GRCS) component in order to demonstrate that we reach state-of-the art results with our settings and to show that the performance differences are related to speaking style (and not related to recording condition or speaker identity).

4.2.4.1 Methods

When training or fine-tuning an ASR system with speech data from GRASS, in all experiments, we performed leave-p-out cross-validation by measuring WERs for specific test splits while training with remaining training splits. In particular, each test split related to one test speaker in case of RS or one test conversation in case of CS, while ensuring speakers in the test set were completely distinct from those used for training. In case of RS, we compare WERs from 37 speakers, and in case of CS, we compare WERs from 19 conversations. In general, if the AM of a proposed system was trained (or fine-tuned) with a given training split, in case of RS only the RS component and the command component from GRASS was used and in case of CS only the CS component of GRASS was used (cf. Sec. 4.2.3). Furthermore, in case of the RS experiments when decoding a specific speaker test split, we excluded the data from the commands component before scoring – in the sense of calculating the WERs – by reporting only WERs for the RS component.

ASR with Whisper (zero-shot): We transcribed Austrian German RS and CS with the same Whisper model (large-v2) (OpenAI, 2023) by setting the parameter `language` to `German`. Furthermore, we set the parameter `suppress_tokens` to -1 and the parameter `temperature_increment_on_fallback` to `None` in order to ensure the suppression of most special characters and the generation of a deterministic output. All other parameters had their default values. Radford et al. (2023) recommended a specific text standardization of the output transcriptions for non-English text, but we chose our own standardization since this improved our results for all conversations. This means that we removed all punctuation (including i.a. brackets), standardized to lowercase and transformed all numbers to German words. Additionally, we standardized typical backchannels (`mh, hm, mmh, hhm, uh huh`) to `mhm`.

ASR with Kaldi: In case of ASR for RS with Kaldi, we chose a similar approach as described in Linke, Wepner, et al. (2023). To make the Kaldi experiments comparable to those of the other ASR systems of this study, we changed the mentioned recipe in two ways: First, we used a leave-p-out cross-validation with $p = 1$ describing one test speaker and, second, we only applied two rules to minimize the phone set (R1: replacement to devoice alveolar and postalveloar fricatives and affricatives; R2: splitting of diphthongs) of the standard Austrian German pronunciation (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014).

In case of ASR for CS with Kaldi, we build on a recipe earlier described in Linke, Wepner, et al. (2023). We achieved the best conversation-dependent WERs with Kaldi by using a pronunciation lexicon with the most likely pronunciations per word (average $1.37 - 1.43$ variants per word), and a LM including GRASS CS and additional text from Austrian German subtitles of broadcasts for the deaf and hard of hearing of an Austrian public television service (approx. 220 k sentences) (Linke, Wepner, et al., 2023; *ORF-TVthek: Broadcasts for the Deaf and Hard of Hearing*, n.d.). Apart from that, for this work, the AM was trained entirely with speech data from GRASS CS but in this case, we performed no LM-rescoring with a four-gram which was trained on 5M German sentences.

The major difference, however, lay in the training of the hybrid model: Given the final GMM-HMM system in this case, we trained a DNN-HMM hybrid model based on another Kaldi recipe (Povey et al., 2022) which uses the chain2 component of the Kaldi toolkit (Povey et al., 2011). This component adopts the LF-MMI criterion (Povey et al., 2016) by computing posteriors from the numerator graph and the denominator graph (cf. equation (4.7) in Sec. 4.1.2). Furthermore, we trained with speed-pertubed 3-fold augmented data (Ko et al., 2015), 40-dimensional high resolution MFCCs$+\Delta+\Delta\Delta$ and 100-dimensional i-vectors in order to perform instantaneous adaption of the neural network (Saon et al., 2013). The network included 12 TDNN-F layers (Povey et al., 2018) with time strides (`1, 1, 1, 0, 3, 3, 3, 3, 3, 3, 3, 3`) where a time stride of `3` for a particular layer means that it comprises three time steps to the left, the central time step, and three time steps to the right given the preceding layer. We trained the model with the natural gradient SGD optimizer and set the variable mini-batch size to `128,64`. We applied a dynamic dropout rate starting with a dropout of 0, increasing to 0.5 at the 50% mark of the training, and then returning to a value of 0. We applied a learning rate schedule which started at an initial learning rate of 0.0005 and decayed during training to a value

of 0.00005. The frame-subsampling factor is set to 3 leading to an output frame rate which is one-third of the regular frame rate. This factor allows some kind of data augmentation since different versions of the training data can be generated by shifting the frames by 0, 1, or 2 frames (this is done "on the fly" in Kaldi's chain models). Each model was trained with a GeForce GTX 1080 Ti GPU which provides 11GB of RAM.

ASR with wav2vec2: With the wav2vec2 framework, we conducted two experiments that are based (cf. Sec. 4.1.2) on GRASS CS and similar pre-processing steps, where one experiment was in lexicon-free mode and one incorporated a lexicon and LM. In general, we trained character-based wav2vec2 models and for the experiment with lexicon and LM, we utilized a simple character-based lexicon where each word maps directly to a character sequence and a LM which was trained uniquely with data from GRASS CS (LMs of order 3 were trained with modified Kneser-Ney smoothing and default pruning by utilizing the KenLM toolkit (Heafield, 2011)). For both experiments, we fine-tuned the pre-trained XLSR model (Facebook Research, 2022) which was trained on 56 000 h of multilingual speech data (Conneau et al., 2021) with a CTC loss (Graves, Fernandez, et al., 2006) (cf. Sec. 4.1.2) by utilizing the fairseq toolkit (Ott et al., 2019). During fine-tuning, the initial learning rate of the model was set to 0.00003 accompanied by a tri-stage learning rate scheduler which divided the training process into three phases with ratios of 0.1, 0.4, and 0.5, culminating in the final phase at 5% of the initial learning rate. The updates for the multi-layer convolutional feature encoder were disabled during fine-tuning. Optimization was achieved using the Adam optimizer and each model was trained with GPUs which provide at least 11GB of RAM due to constrained GPU resources in our laboratory.

4.2.4.2 Results

We present WERs for read speech (RS), separately for each speaker (cf. Tab. 4.3) and separately for each conversation for the CS component (cf. Tab. 4.3 and Figs 4.2 and 4.1). Tab. 4.3 additionally shows the conditions for each ASR architecture in order to clarify if pre-training (PT), fine-tuning (FT), a lexicon (Lex) or a LM was involved.

Speaker-dependent WERs for read speech: For the RS component, we achieved mean speaker-dependent WERs of 11.8% (Whisper), 3.62% (Kaldi), 1.81% (w2v) and 1.01% (w2vLM) with corresponding standard deviations of 2.77% (Whisper), 3.02% (Kaldi), 2.21% (w2v) and 1.61% (w2vLM). Although the absolute ranges between best and worst WERs exceeded approx. 9% (Whisper), 11% (Kaldi), 10% (w2v) or 7% (w2vLM), the small standard deviations ($\leqslant 3.02\%$) suggest that the worst WERs might be outliers. This observation is further supported by the fact that only two speakers had comparatively high WERs in case of Whisper (16.19% and 16.03%), but only one speaker had the worst WERs with all other architectures (cf. Tab. 4.3).

Connection to Speaking Style: Across all ASR architectures, standard deviations of WERs were lower than approx. 3%. This observation may indicate

Table 4.3: Results for speaker-dependent WERs (RS) and conversation-dependent WERs (CS) coming from different ASR architectures. The table displays the WERs [%] for each architecture derived from different conditions with respect to pre-training (**PT**), fine-tuning (**FT**), utilized lexicon (**Lex**), and utilized language model (**LM**). The best and worst WERs, as well as the means and standard deviations of WERs ($\mu \pm \sigma$) are also shown. The Whisper architecture was only pre-trained (zero-shot), while the Kaldi architecture was not pre-trained at all but trained entirely on low-resourced GRASS (either RS or CS). In case of RS, the Kaldi architecture included a phonetic Austrian German standard lexicon$^+$ where two rules were applied in order to minimize the phone set (cf. Sec. 4.2.4). In case of CS, the Kaldi architecture included an advanced phonetic Austrian German pronunciation lexicon* containing most likely pronunciations and a 3-gram LM trained with additional Austrian German subtitles*. Decoding of the wav2vec2 architecture was done by using two methods: lexicon-free decoding (w2v) and decoding with both a simple character-based lexicon and a 3-gram LM (w2vLM).

ASR architecture	**Conditions**				**WERs (RS)**			**WERs (CS)**		
	PT	**FT**	**Lex**	**LM**	**Best**	**Worst**	$\mu \pm \sigma$	**Best**	**Worst**	$\mu \pm \sigma$
Whisper (zero-shot)	✓	–	–	–	6.78	16.19	11.8 ± 2.77	26.45	63.83	41.78 ± 8.23
Kaldi	–	✓	✓$^{+*}$	✓*	0.67	12.4	3.62 ± 3.02	33.06	51.58	42.86 ± 4.78
w2v	✓	✓	–	–	0.15	10.60	1.81 ± 2.21	20.89	38.67	29.81 ± 4.80
w2vLM	✓	✓	✓	✓	0.00	7.67	1.01 ± 1.61	15.27	30.47	22.79 ± 4.02

that all ASR architectures are quite robust to variety-specific pronunciation for Austrian German read speech, or that most Austrian speakers read close to the German standard (when leaving aside the few outliers). Note that in read speech, the Austrian pronunciation is only different to the standard spoken in Germany with respect to a relatively small set of segmental acoustic characteristics (e.g., devoicing of alveolar fricatives, aspiration of plosives, small shifts in the vowel space). Hence, achieving robust state-of-the-art ASR results on read speech, even of the low-resourced variety Austrian German, appears feasible with all four ASR systems presented here **(Relates to A2)**.

Connection to ASR Technology: For RS, we achieved worst WERs with Whisper ($\mu = 11.8\%$) but better WERs with Kaldi and wav2vec2 ($\mu \leqslant 3.84\%$). This indicates that the zero-shot approach with Whisper does not generalize sufficiently well to the Austrian German variety in case of RS **(Relates to A3)**. On the other hand, the low-resourced Kaldi system (≈ 4.6 h of training data per speaker) performs almost as good as a fine-tuned wav2vec2 architecture (XLSR was pre-trained on 56 000 h of multilingual speech data). This demonstrates that achieving state-of-the-art ASR results for the RS component of GRASS is also possible with much less speech data **(Relates to A3)**.

Conversation-dependent WERs for conversational speech: For the CS component, we achieved mean conversation-dependent WERs (cf. Tab. 4.3 and Fig. 4.1) of 41.78% (Whisper), 42.86% (Kaldi), 29.81% (w2v) and 22.79% (w2vLM) with corresponding standard deviations of 8.23% (Whisper), 4.78% (Kaldi), 4.8% (w2v) and 4.02% (w2vLM). The best WER of 15.27% was achieved with w2vLM and the worst WER of 63.84% was achieved with Whisper.

Fig. 4.2 shows normalized histograms of the conversation-dependent WER. We observe a higher standard deviation for Whisper (red: 8.23%) and lower standard deviations for Kaldi, w2v and w2vLM ($\leqslant 4.8\%$). Note that the coefficients of variation for Whisper (19.7%), w2v (16.1%) and w2vLM (17.6%) were higher than

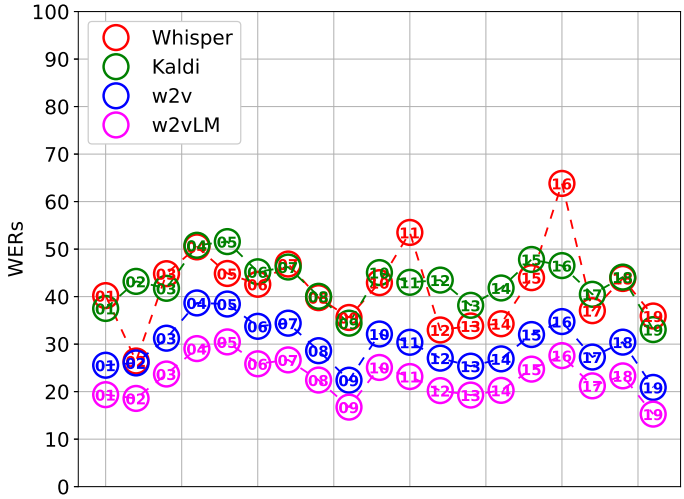


Figure 4.1: Conversation-dependent WERs with respect to the 3 ASR architectures Whisper, Kaldi and wav2vec2 with and without lexicon/LM (w2v and w2vLM). Numbers in circles indicate conversation IDs from the 19 GRASS conversations.

the coefficient of variation for Kaldi (11.2%). Fig. 4.1 shows absolute WERs with respect to the 19 conversation IDs. Interestingly, the same conversation achieved the best WER across the ASR architectures Kaldi and wav2vec2 (cf. conversation ID 19). As it can be further seen in Fig. 4.1, WERs obtained by different ASR architectures seem to be strongly correlated, i.e., have high Pearson correlation coefficients (cf. Fig. 4.3a) of 60.4% (Kaldi, Whisper), 89.8% (Kaldi, w2v), 87.2% (Kaldi, w2vLM), 80.2% (Whisper, w2v), 81.9% (Whisper, w2vLM), and 99.1% (w2v, w2vLM). The corresponding ranking of the conversation-dependent WERs of each ASR system allows the calculation of Spearman rank-order correlation coefficients (cf. Fig. 4.3b) which were high in case of w2v and w2vLM (75.8%) but negative and close zero in case of Kaldi and Whisper (−6.1%). Simultaneously, the rank-order correlation coefficient was moderately high and positive in case of Kaldi and w2vLM (40.4%) but moderately low and negative in case of Whisper and w2v (−21.4%) or Whisper and w2vLM (−18.4%). That Whisper has small or even negative rank correlations with other ASR systems while still having strong positive Pearson correlations can be explained by the fact that rankings of Whisper often disagree with rankings of other ASR systems despite showing similar general trends (e.g., compare the relative ranks of conversations 12–14 or 3–5). Additionally, the rank-order correlation coefficient between Kaldi and w2v was lower than the rank-order correlation coefficient between Kaldi and w2vLM (21.1% < 40.4%) which we believe can be attributed to the fact that both Kaldi and w2vLM make use of a LM.

Connection to Speaking Style: We achieved high standard deviations (> 4%) for conversation-dependent WERs across all ASR architectures, with Whisper having the highest standard deviation of 8.23%. Simultaneously, Pearson correlation

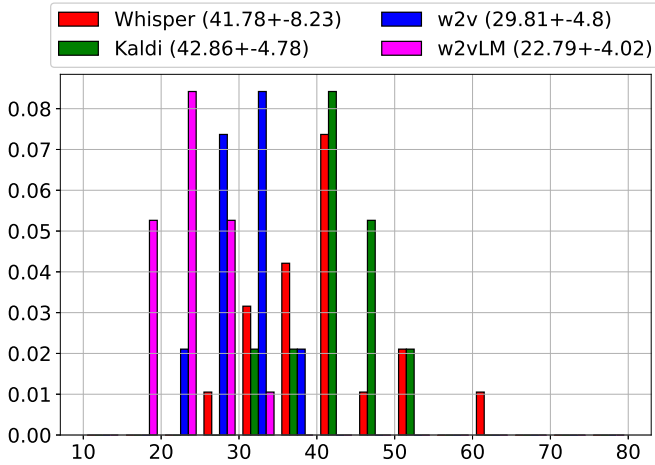
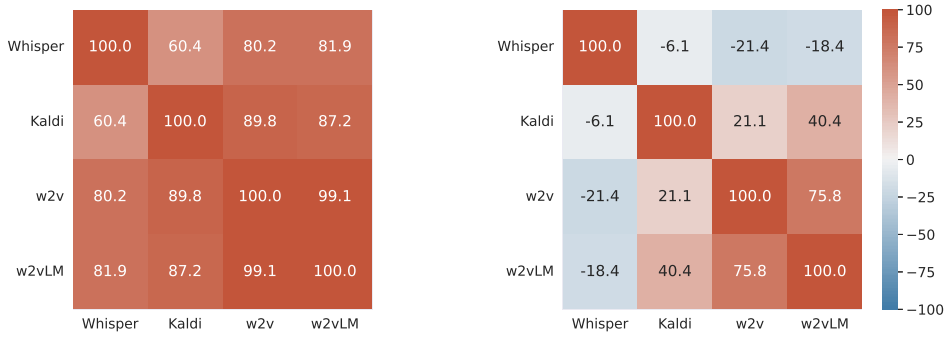


Figure 4.2: Normalized histograms estimating the probability density functions of the conversation-dependent WERs [%] coming from the 3 ASR architectures Whisper, Kaldi and wav2vec2 with and without lexicon/LM (w2v and w2vLM).

coefficients were high ($> 60\%$ in case of Whisper versus all and $> 87\%$ in case of Kaldi versus wav2vec2 architecture) but Spearman rank-order correlation coefficients were generally lower especially with respect to Whisper and Kaldi[1] ($< |22\%|$ in case of Whisper versus all, $< |41\%|$ in the case of Kaldi versus all). With respect to the coefficient of variation, Kaldi had the smallest value ($11.2\% < 16.1\%$), while the values of the other architectures were similarly high ($\geqslant 16.1\%$). The presence of a strong linear relationship (Pearson) coupled with a weak monotonic relationship (Spearman) between the conversation-dependent WERs demonstrates a robustness problem in case of conversational speech recognition with Austrian German which was particularly evident in the case of Whisper. Simultaneously, the coefficient of variation for Kaldi indicates more robustness in comparison to the other architectures. Nevertheless, in general, we found a complex variability in ASR performance across all architectures which highlights the challenge of achieving actual robust ASR results with current state-of-the-art ASR architectures. Given the overall high correlation across ASR systems of which conversations were best or worst recognized, we conclude that the performance variation is related to conversation-intrinsic characteristics. The analysis in Sec. 4.2.6 aims at investigating which features capturing conversation-dependent variation best explain the variation observed across the four ASR systems.

Connection to ASR Technology: For Whisper and Kaldi, we achieved worst conversation-dependent WERs with means of 41.78% and 42.86%. However, with the wav2vec2 architecture, we achieved best results with mean conversation-dependent WERs lower than 30%. These results demonstrate that a zero-shot ASR system (Whisper) which was trained on enormous amounts of multilingual (out-of-domain) speech data (680 000 h) and a low-resourced ASR system (Kaldi), which was trained

[1]The wav2vec2 architecture constitutes an exceptional case since the only difference between the results from w2v and w2vLM was the decoding strategy.



(a) Pearson correlation coefficients. (b) Spearman rank-order correlation coefficients.

Figure 4.3: Comparison of linear (a) and monotonic (b) relationships [%] between the 3 ASR architectures Whisper, Kaldi, wav2vec2 with and without lexicon/LM (w2v and w2vLM). The color bar refers to both matrices.

entirely on (in-domain) speech data (a share of approx. $\frac{13.5\,\text{h}}{680\,000\,\text{h}} \hat{=} 0.002\%$), both achieve poor performance for Austrian German CS **(Relates to A2)**. Simultaneously, fine-tuning the wav2vec2 architecure (pre-trained on 56 000 h of multilingual speech data) with (in-domain) speech data (a share of approx. $\frac{13.5\,\text{h}}{56\,000\,\text{h}} \hat{=} 0.025\%$) and decoding with a lexicon/LM improved the mean of the conversation-dependent WERs by approx. 20% **(Relates to A3)**. Summing up, our results for Austrian German CS indicate that state-of-the-art ASR architectures fail to achieve satisfactory performance, while showing the benefits of pre-training on a substantial amount of speech data, subsequent fine-tuning and providing additional linguistic variety and style specific knowledge (i.e., a LM and a lexicon for Austrian German)[2].

4.2.5 Acoustic and lexical feature extraction

We extracted 12 features on utterance level in order to evaluate their relationship to the observed WERs achieved with each of the four ASR systems. The choice to extract features at the utterance- and not at the word level was motivated by the study of Hirschberg et al. (2004), which aimed at predicting WERs for turns in human-machine interaction. The 12 extracted features are related to utterance length (3), prosody (5), pronunciation variation (2) and perplexity (2). This section describes in detail how these features were calculated and shows an analysis of how strongly these features correlate in general with the conversation-dependent WERs and WERs on utterance level coming from each ASR architecture.

4.2.5.1 Utterance length features

Motivation: In earlier WER analyses on HMM-based systems, it has been shown that longer turns (as measured in seconds) in human-machine interaction have on average more WERs than shorter turns (Hirschberg et al., 2004). We assume that

[2]We assume that fine-tuning Whisper has the potential to yield additional enhancements in speech recognition performance.

this effect will also be seen with respect to utterance-level duration in human-human conversation, at least with respect to the results from the Kaldi-based ASR system.

Calculation: We calculated three different utterance-length features: The first feature is the number of word tokens per utterance (**#tokens**), calculated by counting the words in the pre-processed reference transcriptions (cf. Sec. 4.2.3). The second utterance length feature is based on a forced alignment (FA) and counts the number of realized phones per utterance (**#phones**). For the Kaldi-based forced alignment, a lexicon with multiple pronunciation variants per word was used, where these pronunciation variants included specific variants typical for Austrian German casual spontaneous speech (for more details cf. Linke, Wepner, et al. (2023)). The third feature (**UttDur**) is the total duration of each utterance measured in seconds (including potential short silences and speaker noises such as breathings etc.).

4.2.5.2 Prosodic features

Motivation: Earlier analyses of the performance of HMM-based systems on spontaneous speech showed that prosodic characteristics correlate with WERs (Goldwater et al., 2008; Hirschberg et al., 2004). As in these studies, we consider articulation rate and F0- and RMS-related features. Whereas these earlier studies considered the mean, max, min and range of F0 and RMS, we decided to use entropy measures in order to capture the total variation of F0 and RMS over the utterance.

Calculation: The first prosodic feature is the average articulation rate (**AR**) over the utterance, i.e.,
AR = **#phones**/**UttDur**. A comparable speech rate feature on word level has earlier been reported to be a strong predictor for WER (Goldwater et al., 2008). We calculated the articulation rate by dividing the number of realized phones as given by the forced alignment (without the silence phones) by the duration of the utterance. It thus is an articulation rate in the strict sense, not to be confused with the local speech rate (i.e., including silence durations into the measure).

Second, we calculated F0 and RMS related features with a similar approach described in Linke, Kubin, and Schuppler (2023). For both F0 and RMS extraction at utterance level, we used *pyreaper*, where for F0, we used the default settings (Google, 2023; Yamamoto, 2023) and for RMS extraction, we defined frames with a frame length of 40 ms and a frame shift of 10 ms. After that, we calculated two types of entropy measurements for both contours. In general, entropy is a measure of the spread of probability distributions and reflects the uncertainty associated with a random variable X (Cover & Thomas, 2006). When the random variable X takes on values $x_i \in \mathcal{X}$, where $\mathcal{X}$ is a finite set, its entropy is defined as can be expressed as

$$H(X) = -\sum_i p_i \cdot \log p_i, \tag{4.12}$$

where $p_i = Pr\{X = x_i\}$ is the probability of X taking the value x_i and where $0 \log 0 := 0$ by convention.

Given a sequence of N non-negative feature values $\langle f[1], f[2], \ldots, f[N] \rangle$ observed within an utterance (such as F0 or RMS values), we can quantify the spread of these values by **(a)** estimating a probability distribution via a histogram leading to the

conventional entropy measure or **(b)** estimating a (pseudo-)probability distribution leading to a formal (pseudo-)entropy measure. Hence, in the first case **(a)**, we generated normalized histograms by binning the feature values between 80 Hz and 240 Hz with a width of 5 Hz (F0 contours) or between 0 and 1 with a width of 0.05 (RMS contours) which estimates *conventional* probability distributions. Likewise, in the second case **(b)**, we estimated (pseudo-)probability distributions which are defined by normalizing the feature values

$$p_i = \frac{f[i]}{\sum_{n=1}^{N} f[n]} \tag{4.13}$$

and ensuring that the total probability condition $\sum_{i=1}^{N} p_i = 1$ is satisfied. Using this second approach, the entropy (4.12) attains its maximum value $H_{max} = \log N$ when the feature sequence is constant, i.e., when $f[1] = f[2] = \cdots = f[N] = \text{const}$. On the other hand, the minimum value of entropy $H_{min} = 0$ is achieved when the probabilities in equation (4.13) are close to either 1 or 0, which can occur in the case of a highly non-uniform feature sequence within a utterance. It is worth noting that this (pseudo-)entropy measure quantifies the (relative) variability of the features within the utterance, without taking into account the time order of the feature contour.

Finally, we normalized both entropy measures **(a)** and **(b)** by dividing the resulting entropies by the logarithm of the number of bins (in case of the *conventional* entropies) or the logarithm of the sequence length[3] (in case of the (pseudo-)entropies):

$$\tilde{H} = \frac{H}{\log N}. \tag{4.14}$$

This final normalization results in two F0/RMS features represented as *conventional* entropies **HF0N** as well as **HRMSN** and two F0/RMS features represented as (pseudo-)entropies **HPSF0N** and **HPSRMSN**.

4.2.5.3 Pronunciation features

Motivation: Earlier analyses on WERs have rather considered pronunication-related features motivated by psycholinguistic findings indicating that human subjects have more difficulty recognizing spoken words that are in dense phonetic neighborhoods (Goldwater et al., 2008). These works, however, did not deal with speech from a low-resourced variety of a language containing dialectal data as we do. For our purposes, we thus chose to extract features that reflect how strongly the pronunciation of an utterance differs from the canonical, standard pronunciation of the word sequence spoken. Given that most available German speech material used to train the models contains speech that is either prepared or not Austrian, we assume that utterances closer to standard pronunciation are better recognized. We extracted two pronunciation features that reflect the degree of acoustic reduction (which has been shown to increase with the degree of spontaneity (Adda-Decker & Lamel, 2018)), and the extent of dialectal pronunciation.

[3] In case of F0 contours, we calculated the sequence length N by excluding unvoiced segments.

Calculation: We extract two features, where the first one, **PronD**, reflects the degree of reduction of the utterance, and the second one, **PronLD**, reflects both the degree of reduction and deviation from the standard pronunciation. In more detail, **PronD** was calculated as the mean of the differences between the number of phones of the not-reduced, canonical pronunciation[4] of a word and the number of phones of the actually realized pronunciation of the same word. Hence, the resolution of this feature increases with the number of phones per word and the number of words per utterance. For the calculation of the second pronunciation feature **PronLD**, instead of the mean of ordinary differences, we measured the mean of the Levenshtein distances (Levenshtein, 1965) between the realized pronunciation of a word to its canonical pronunciation in the lexicon for standard Austrian German. The second pronunciation feature **PronLD** is more comprehensive as it considers not only deletions but also substitutions and insertions.

4.2.5.4 Perplexity features

Motivation: In casual, spontaneous interaction, utterances are often short, fragments of sentences, containing repetitions, disfluencies and, as we deal with Austrian German, dialectal grammatical structures, which all would not be found in written text nor in prepared speech. In order to measure how strongly the word sequence of an utterance in the GRASS reference transcriptions 'comes as a surprise' to a LM, we calculated two LM perplexity features, one based on written German, and one based on spoken Austrian German (inspired by Goldwater et al. (2008), who analzed WERs with respect to the trigram-log-probability). We assume that ASR architectures that were specifically trained on (even if only small amounts of) data from conversational speech can better deal with small, fragment-style utterances in GRASS.

Calculation: We provide two perplexity features **pplAGS** and **pplWIKI**. As earlier described in Linke, Wepner, et al. (2023), we measured perplexities with a trigram LM (**pplAGS**) trained on 220 k sentences from an Austrian German television service (*ORF-TVthek: Broadcasts for the Deaf and Hard of Hearing*, n.d.) and a four-gram LM (**pplWIKI**) trained on a subset of 5 M German sentences which originated mainly from German Wikipedia and the European parliament. In general, perplexities were calculated with the SRILM toolkit (Stolcke, 2002) which incorporate sentence-beginning (`S_B`) and sentence-ending (`S_E`) markers for the perplexity calculation to fully capture the contextual boundaries of each utterance. This means that the probability of the first token of an utterance is given by the probability of the first word w_1 leading to $P(w_1|\texttt{S\_B})$. Simultaneously, the probability of the last token would always refer to the probability of the sentence-ending. For instance, in case of an utterance including only one word token w_1 this probability would be $P(\texttt{S\_E}|w_1, \texttt{S\_B})$. Thus, in case of an utterance including two word tokens w_1 and w_2 the probability would be $P(\texttt{S\_E}|w_2, w_1, \texttt{S\_B})$. The perplexity of an utterance is related to the negative sum of the logarithms of the conditional

[4]This canonical pronunciation is derived from a pronunciation lexicon for standard Austrian German, cf. Linke, Wepner, et al. (2023).

probabilities via

$$\log_{10} \mathbf{ppl} = -\frac{1}{\#\mathbf{tokens}+1} \sum_{l=1}^{\#\mathbf{tokens}} \log_{10} P(w_l | w_{l-1}, \ldots, w_{l-n+1}) \tag{4.15}$$

where n is the order of the n-gram model, $w_{\leqslant 0}$ is `S_B` and $w_{\#\mathbf{tokens}+1}$ is `S_E`, and where $P(\cdot)$ is given by the LM. Note that the perplexity is normalized by the utterance length while the probability for `S_E` was also considered. Eq. 4.15 was used to measure perplexities **pplAGS** (order $n = 3$) and **pplWIKI** (order $n = 4$) for each utterance.

Summing up, we extracted 12 features, falling into the following categories:

1. Utterance length features (3): **#tokens**, **#phones** and **UttDur**.
2. Prosodic features (5): 1 durational feature: **AR**, 2 F0 features: **HF0N** and **HPSF0N**, 2 RMS features: **HRMSN** and **HPSRMSN**
3. Pronunciation features (2): **PronD** and **PronLD**
4. Perplexity features (2): **pplAGS** and **pplWIKI**.

4.2.6 Analysis: How do acoustic and lexical utterance features affect the performance of different ASR systems?

We analyze the relationships between the WERs of GRASS CS on utterance level (the *dependent variable*) with respect to each ASR architecture and specific features (the *independent variables*) from the different feature categories (cf. Sec. 4.2.5) by performing a statistical analysis.

The remainder of this section is structured as follows: First, we motivate our methodological approach in Sec. 4.2.6.1 by describing the distribution of the dependent variable, namely the WERs on utterance level, separately for each of the four ASR systems and as a function of the utterance length. Second, we describe how the extracted features (i.e., the features capturing spontaneous speech phenomena from Sec. 4.2.5) correlate with the WERs on utterance level, and third, we describe the feature selection approach for the independent variables (cf. Sec. 4.2.6.2). Finally, we present our statistical analysis by means of *Interaction Forests* (Hornung & Boulesteix, 2022a) in Sec. 4.2.7, where we separately discuss the results from the importance measurements for univariable effects and quantitative and qualitative interactions.

In the following, utterances that contain only one word token are referred to as *single-word* utterances, those that are between one and four word tokens long are referred to as *short* utterances and those that are between 5 and 15 word tokens long are referred to as *long* utterances.

4.2.6.1 How do utterance lengths affect WERs on utterance level?

To get a global picture of how the WERs on utterance level (the dependent variable of our statistical analysis) were distributed with respect to each ASR architecture, we computed histograms separately for utterances of different length (cf. Fig. 4.4). This was motivated by the fact that, in contrast with other containing less spontaneous

or less interactional speech, the GRASS CS corpus contains a high number of *single-word* utterances (cf. Fig. 4.4a).

When focusing on the distribution of WERs for *short* utterances (cf. Fig. 4.4e) we observe a large difference to the distribution for *long* utterances (cf. Fig. 4.4f). For *short* utterances, the majority of WERs were at 0% or 100%. With respect to the different ASR architectures, in case of Whisper and Kaldi approx. 10000/12500 utterances led to a WER of 0% and approx. 7500/5000 utterances led to a WER of 100%. In contrast, in case of the wav2vec2 architecture (i.e., w2v and w2vLM), approx. 15000/17500 utterances led to a WER of 0% and approx. 2500/2500 utterances led to a WER of 100%. Note that WERs of 90% did not (really) occur for any ASR system, a phenomenon for which we do not have any data-related nor methodological explanation.

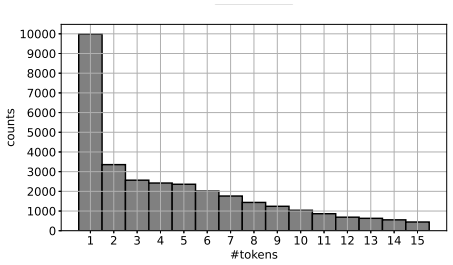
To determine how often WERs of 0% and 100% resulted from *short* utterances, we show a separate histogram for this subset of the corpus in Fig. 4.4c. It can be seen that of 10000 utterances with a single word, Whisper correctly recognizes $\approx$ 6000 utterances, while the other architectures correctly recognize more than $\approx$ 8000. This shows a remarkable difference between Whisper (zero-shot) and the other ASR systems that have domain-knowledge.

The histogram of WERs for utterances containing two to four word tokens (cf. Fig. 4.4d) shows a different behaviour with respect to the WERs at 0% and 100%. Now, approx. 2000/3000/4000/5000 (Whisper/Kaldi/w2v/w2vLM) utterances had WERs of 0% and approx. 3000/2000/1000/500 (Whisper/Kaldi/w2v/w2vLM) utterances had WERs of 100%. For utterances of these lengths, the counts were more or less similar across all ASR architectures for WERs of 20%, 30%, 60% and 70%.

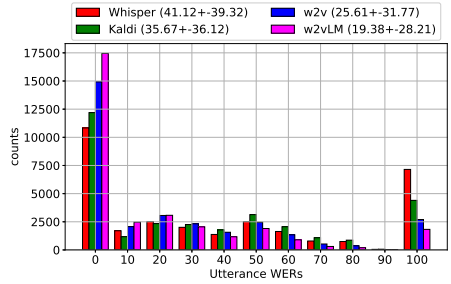
Finally, we also present WERs on utterance level with respect to the data used in the two experiments of our statistical analysis in Fig. 4.4e and 4.4f. In general, the histogram in Fig. 4.4e comprises the data of the histograms in Fig. 4.4c and 4.4d. We observe a decreasing trend with respect to the WERs when analyzing the WERs of *long* utterances (cf. Fig. 4.4f). In general, we find that Kaldi achieved worse WERs for longer utterances (i.e., of more than four word tokens).

Fig. 4.5 gives another representation of the WERs on utterance level which also summarizes specific phenomena explained by the histograms 4.4b-4.4f in Fig. 4.4. We plotted the mean WERs on utterance level for specific numbers of word tokens within an utterance across all ASR architectures. This illustration shows that Whisper achieved worse WERs for *short* utterances (mean WER of 43.7%) than for *long* utterances (mean WER of 37.5%), while for all other systems this direction was reversed ($30.2\% < 43.5\%$, $22.4\% < 30.2\%$ and $16.8\% < 23\%$). Simultaneously, we also observe that Kaldi and the wav2vec2 architecture seem to be quite robust for utterances containing $2 \ldots 15$ word tokens since mean WERs remained almost constant (approx. 45%/30%/25% in case of Kaldi/w2v/w2vLM). There was only one slight exception in case of Kaldi because for utterances containing two word tokens the mean WER was slighly lower ($\approx$ 40%). Finally, we report a relevant observation with respect to the wav2vec2 architecture: The inclusion of linguistic knowledge (i.e, decoding with a lexicon and LM) led to a substantial improvement of the mean WER by $\approx$ 5.6% (for *short* utterances) and $\approx$ 7% (for *long* utterances).

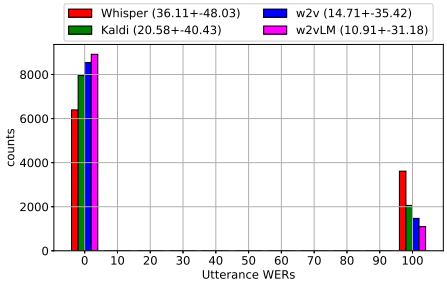
Connection to Speaking Style: We observe that the GRASS CS corpus shows a decreasing trend with respect to the counts of word tokens per utterance ($\approx$ 10000



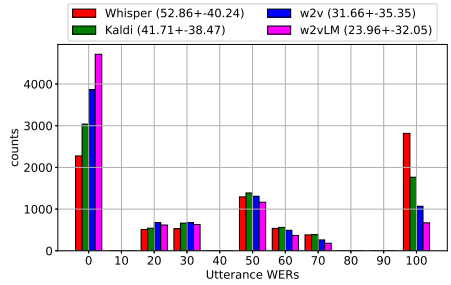
(a) counts of utterances with respect to the number of word tokens



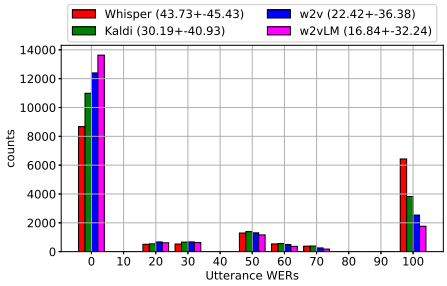
(b) 1 . . . 15 word tokens



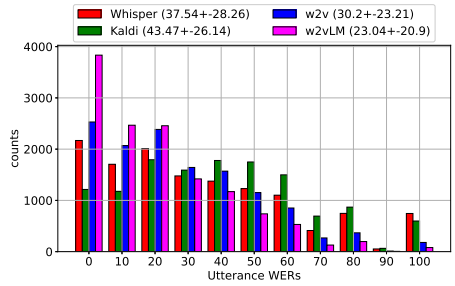
(c) one word token



(d) 2 . . . 4 word tokens



(e) 1 . . . 4 word tokens



(f) 5 . . . 15 word tokens

Figure 4.4: Histograms describing the utterances of GRASS CS. The upper left histogram (4.4a) shows the counts of utterances per number of word tokens. All other histograms show the counts of WERs [%] on utterance level with respect to the 3 ASR architectures Whisper, Kaldi and wav2vec2 with and without lexicon/LM (w2v and w2vLM). In particular, these histograms refer to utterances containing 1 . . . 15 word tokens (4.4b), one word token (4.4c), 2 . . . 4 word tokens (4.4d), 1 . . . 4 word tokens (4.4e and 5 . . . 15 word tokens (4.4f) where legends include respective mean WERs on utterance level and corresponding standard deviations (i.e., $\mu \pm \sigma$).

utterances containing one word token in contrast to less than 450 utterances containing 15 word tokens). As a consequence, ASR performance for CS is strongly affected by utterances with less than five word tokens (in case of the zero-shot Whisper architecture the mean WER improved for utterances containing more word tokens, potentially due to its transformer architecture exploiting the context better, and in case of the trained or fine-tuned architectures the mean WER improved for utterances containing less word tokens which was mainly explained by a better performance for *single-word* utterances). Our results indicate that the speaking

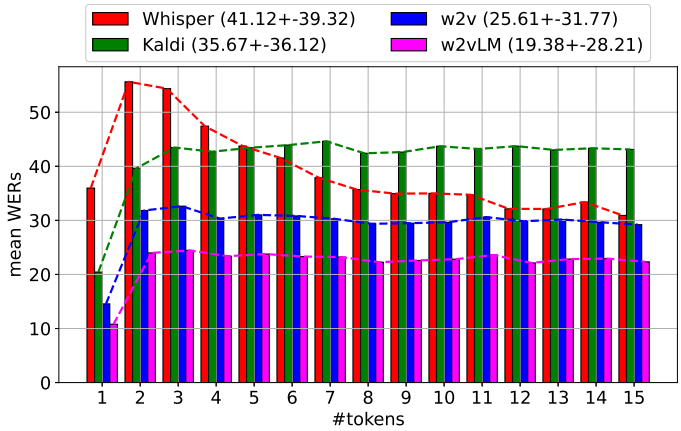


Figure 4.5: Mean WERs [%] on utterance level for specific numbers of word tokens within an utterance across all ASR architectures. The legend shows again the overall mean and standard deviation of WERs on utterance level across all 3 ASR architectures Whisper, Kaldi and wav2vec2 with and without lexicon/LM (w2v and w2vLM)

style CS is characterized by utterances containing less than ≈ 5 word tokens (e.g., backchannels, short response tokens, broken phrases, etc.), a characteristic resulting from spontaneous speech where two speakers are in constant interaction. Our results suggest that for conversational speech, ASR performance may be improved if ASR architectures provide learning strategies to better capture important short-term dependencies (with short-term dependencies, we mean here mainly the dependencies with respect to only one, two, three or four word tokens). We also recognize the benefit of learning the dependencies in two steps, which is basically the concept of the wav2vec2 architecture **(Relates to A3)**. This architecture is also based on a context network but in a first stage, it was fine-tuned based solely on a CTC loss for character sequences (short-term dependencies (bottom-up)) and in a second stage, it was decoded with a lexicon and LM (long-term dependencies (top-down)).

Connection to ASR Technology: Only in case of Whisper mean WERs on utterance level were better for less word tokens per utterance ($\approx 43.7\%$) and worse for more word tokens per utterance ($\approx 37.5\%$). A comparison of w2v and w2vLM demonstrated that the wav2vec2 architecture benefits from linguistic knowledge provided by a lexicon and LM because mean WERs on utterance level were approx. $5\% - 7\%$ better independent of the number of word tokens per utterance. In general, we found that mean WERs for *single-word* utterances were much better than all other mean WERs of utterances containing more word tokens across all ASR architectures (cf. Fig. 4.5). We can state that all trained or fine-tuned ASR architectures were better able to adapt to the speaking style CS which is particularly noticeable in the mean WERs for fewer word tokens. At the same time, however, we must also note that optimal robustness would strictly mean that these better mean WERs must also be available for utterances that contain more word tokens but this is not the case in any analyzed ASR architecture. We can again note that Whisper behaves differently than the trained or fine-tuned architectures as the ASR performance

of Whisper improves with an increased number of word tokens. This may be an explanation for why Whisper showed higher standard deviations for the mean conversation-dependent WER. Importantly, regardless of the number of word tokens within an utterance, a constant performance gain was reached by incorporating linguistic knowledge to the wav2vec2 **(Relates to A3)**. To conclude, the analysis of the independent variable reveals that the performance of ASR architectures varies greatly with the number of word tokens per utterance and that the performance of a zero-shot architecture improves especially with an increased number of word tokens. Our results also highlight the importance of linguistic knowledge and the need for further optimization to achieve consistent robustness across varying utterance lengths.

4.2.6.2 How do acoustic and lexical features affect WERs? Correlation analysis and feature selection

This section describes how features on utterance level correlated with WERs with respect to each ASR architecture and how we selected our features (the independent variables) for the statistical analysis.

Feature correlation with the WERs? As a next step in our analysis, we estimate which of the extracted features (cf. Sec. 4.2.5) correlate how strongly with the WERs of the different ASR systems. Fig. 4.6 shows resulting Pearson correlation coefficients between the overall WERs of each of the systems Whisper, Kaldi, wav2vec2 with and without lexicon/LM (w2v and w2vLM) and the extracted utterance-level features. In order to evaluate correlations which are independent of the conversation IDs, we compared the features by correlating the feature values with all WERs on utterance level.

Consistent with what we have discussed in the previous section, we observe weak negative correlations between WERs and utterance length features for Whisper ($> -12\%$), and weak positive correlations for all other systems ($< 18\%$), with those for Kaldi being the strongest. The durational feature **AR** also showed the strongest correlation with the WERs from Kaldi (24.6%) in contrast to Whisper (9.6%). With respect to the F0 features, we observe weak negative (Whisper) or positive (Kaldi and wav2vec2) correlations in case of **HF0N** (between -9% and 12%) but stronger negative correlations in case of **HPSF0N** (Whisper: -16.8%; Kaldi/wav2vec2: between -10% and -12%). This is as expected, as pseudo-entropy is large if the respective contours are flat, while conventional entropy is large if the respective contours span multiple bins used in estimating the probability distribution. Therefore, one would expect that a negative correlation between WERs and pseudo-entropies coincides with positive correlation between WERs and conventional entropies, and vice-versa. In comparison to Kaldi and wav2vec, Whisper's performance was correlated most strongly with **HPSF0N** (-16.8%). All RMS features showed weak negative and positive correlations close to 0% with the exception of Whisper showing negative correlations of -5.6% (**HRMSN**) and -8% (**HPSRMSN**). In contrast, the pronunciation features were strongly positively correlated to the WERs of Kaldi ($\approx 24\%$), w2v ($\approx 21\%$), and w2vLM ($\approx 18\%$). In case of Whisper those correlations were also positive but weaker. Finally, correlations with all perplexity features were positive but weak, and they were weaker in case of Whisper ($< 5\%$) in comparison to the other architectures ($> 5\%$).

Connection to Speaking Style: We observed highest correlations with the durational feature **AR** as well as the pronunciation features **PronD** and **PronLD** across all ASR architectures (mean correlations of highest correlating features with respect to each category $> 55\%/> 19\%$). The F0 feature **HPSF0N** showed slightly higher correlations in case of Kaldi ($-46.4/-10.3$) or Whisper ($-10.4/-16.8$). All of those features represent typical phenomena in complex spontaneous or conversational speaking styles. Especially the pronunciation features reflect one noticeable characteristic of GRASS CS since speakers articulated on average up to $1.37 - 1.43$ variants per word (Linke, Wepner, et al., 2023). We conclude that these features describe common phenomena of the conversational speaking style, which we further analyze in Sec. 5.

Connection to ASR Technology: In case of Whisper, we observe partly contradictory correlations in contrast to the other architectures (especially with respect to utterance duration and F0 entropy). Whisper is the only zero-shot architecture in our analysis which leads to intriguing comparisons with the other trained or fine-tuned architectures since Kaldi and wav2vec2 are regulated by in-domain data. In general, Whisper performs better on *long* utterances than the other architectures, which we explain by its advanced embedding of contextual information.

Feature selection: We selected features for the subsequent analyses based on the correlations described above. In particular, we considered the correlations between all WERs on utterance level (the dependent variable) and all prosodic, pronunciation and perplexity feature values (the independent variables). As the only exception, in case of the utterance length feature, we manually selected **#tokens** because this feature determines the possible resolutions of the WERs which highly affects the distributions of the WERs on utterance level (cf. Fig. 4.4 and Sec. 4.2.6.1). For all other feature categories, in order to mitigate the effects of multicollinearity, we restricted our feature selection to only one feature from each feature category. In case of the durational feature, we selected the only available feature **AR**. In case of the remaining features, we calculated mean absolute correlation values of each feature. This comparison showed that in both cases best prosodic, pronunciation and perplexity features were **HPSF0N**, **HPSRMSN**, **PronLD** and **pplWIKI** with mean correlations of 12.05, 3.94, 20.35 and 6.73. Given that correlations between the entropy features **HPSF0N** and **HPSRMSN** were much higher ($\approx 53\%$ for *short* utterances and $\approx 72\%$ for *long* utterances) than between other prosodic, pronunciation, and perplexity features ($< |28|\%$ and $< |16|\%$ respectively), we addressed the possible issue of multicollinearity in our statistical analysis by decorrelating the feature **HPSRMSN**. Therefore, in order to decorrelate the feature **HPSRMSN** from **HPSF0N**, we replaced **HPSRMSN** with the residuals from the linear model

$$\overline{\mathbf{HPSRMSN}} = \beta_0 + \beta_1 \cdot \mathbf{HPSF0N}. \tag{4.16}$$

This leads to the new feature $\mathbf{HPSRMSN}_{Res}$ which can be calculated as

$$\mathbf{HPSRMSN}_{Res} = \mathbf{HPSRMSN} - \overline{\mathbf{HPSRMSN}}. \tag{4.17}$$

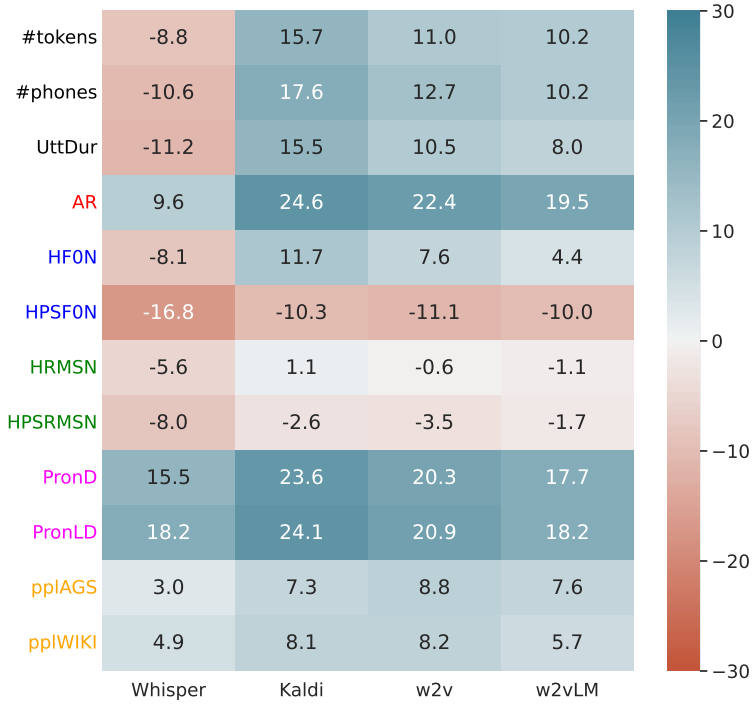


Figure 4.6: Correlations [%] of all feature values with all WERs on utterance level independent of the conversations. The selected features are from six categories: utterance length (black), prosodic (red/blue/green), pronunciation (magenta) and perplexity (orange) features.

These residuals explain the deviation of **HPSRMSN** from the values predicted by the linear model.

As a result, only the utterance length feature **#tokens** was moderately correlated with **PronLD** (37.5%) and **AR** (31.9%) in case of the first statistical analysis (utterances with 1 . . . 4 word tokens) and a weaker correlation with **HPSF0N** (16.2%) in case of the second statistical analysis (utterances with 5 . . . 15 word tokens). Overall, the final feature set for both statistical analyses included the following six features (or independent variables): **#tokens** (utterance length feature), **AR** (durational feature), **HPSF0N** (F0 feature), $\mathbf{HPSRMSN}_{Res}$ (RMS feature represented as residuals given the response **HPSRMSN** and the predictor **HPSF0N**), **PronLD** (pronunciation feature) and **pplWIKI** (perplexity feature).

4.2.7 Statistical analysis with Interaction Forests

This section describes the approach for the statistical analysis. In Sec. 4.2.6.1 we showed that the WERs on utterance level are not normally distributed (cf. Fig. 4.4b) and that the number of available utterances decreases with an increasing number

of word tokens, leading to an insufficient amount of data points (cf. Fig. 4.4a). The comparison of WERs in Sec. 4.2.6.1 also showed how distributions of *short* utterances (cf. Fig. 4.4e) and *long* utterances (cf. Fig. 4.4f) differ. Hence, in order to better capture the relationships between the features and the WERs on utterance level, we restricted our statistical analysis to utterances which have at most 15 word tokens and divided our data in two sets for the analysis:

1. Statistical analysis of *short* utterances (analysis of 18326 utterances).
2. Statistical analysis of *long* utterances (analysis of 13030 utterances).

In general, we based our statistical analysis on *Interaction Forests*, a recently introduced variant of Random Forests (Hornung & Boulesteix, 2022a). Interaction Forests provide *effect importance measures* (EIMs) for univariable effects as well as quantitative and qualitative interaction effects. Thus, in contrast to classical *variable importance measures* (VIMs), EIM values rank not only univariable effects but also effects due to quantitative and qualitative interactions which can be communicated in a comprehensible manner. A short description of these effects can be given as:

1. A univariable effect indicates that the dependent variable changes monotonically with the considered independent variable x. univariable effects are represented by univariate, binary splits in the trees constituting the random forest, e.g., $x < x_s$ versus $x \geqslant x_s$, where x denotes a variable and x_s a split point for this variable x.
2. A quantitative interaction effect indicates that another independent variable y controls the strength with which the dependent variable changes monotonically with the independent variable x. E.g., the dependent variable changes strongly with x if y is small and only weakly if y is large, etc. Quantitative interaction effects are represented by bivariate, binary splits in the trees. E.g., if a dependent variable is particularly small if both x and y are small (but not if they are individually so), then this is represented by a split depending on the truth value of $\{x < x_s\} \cap \{y < y_s\}$. This results in **four split types** corresponding to the four quadrants in the interaction space (x and y small, x and y large, x small and y large, x large and y small).
3. A qualitative interaction effect indicates that the direction of the monotonic effect of the independent variable x on the dependent variable is controlled by another independent variable y. E.g., the dependent variable increases with x for small y but decreases with x for large y. These qualitative interactions are represented by bivariate, binary splits in the trees that are based on the truth value of the statement $\{\{x < x_s\} \cap \{y < y_s\}\} \cup \{\{x > x_s\} \cap \{y > y_s\}\}$.

Fig. 4.7 also visualizes these split types (the visualization was taken from the original paper with a slighlty different notation). EIM values were computed from the resulting Random Forests by first evaluating the accuracy of the constituting trees on the subset of data on which the tree was not trained, and then comparing this accuracy to the one that would be obtained by the same tree if the considered effect (the one univariable, one of the four quantitative interactions, or the one qualitative interaction) is not represented (i.e., when the data is not split according to the learned criterion but randomly assigned to the children of the splitting

node). Since univariable effects can be detected also as quantitative and qualitative interactions (and since splits are generated randomly), the EIM values of quantitative and qualitative interactions are adjusted accordingly. For a more comprehensive understanding of the algorithms associated with Interaction Forests, we refer to the paper by Hornung and Boulesteix (2022a) and the corresponding supplementary material 1 (Hornung & Boulesteix, 2022b).

For each experiment and for each ASR system (which leads to 4 different dependent variables), we trained Interaction Forests in R with the R package `diversityForest` (Hornung & Wright, 2023) and used a default value of 20000 trees which allows a sufficient calculation of the EIM values for the 3 different effect types. All other parameters were also set to their default values.

Hornung and Boulesteix (2022a) claim that p-values should be analyzed with caution because they are generally much too optimistic especially in case of small datasets and large numbers of variables. The cause of this is that tests for interaction effects with classical linear regression lead to p-values which would not be adjusted for the circumstance that the data was already used to find variable pairs which indicated strongest interaction effects (Hornung & Boulesteix, 2022b). However, in our analysis, we considered only six features (or independent variables) which allows a straightforward calculation of Bonferroni-adjusted p-values with only a small number of $\binom{6}{2} = p \cdot 15$ possible interactions. Hence, in our statistical analysis, we report significance levels which are based on Bonferroni-adjusted p-values $p_a = p \cdot \binom{6}{2} = p \cdot 15$ which leads to the following updated significance codes: Three stars (***) indicate that tests for interactions using linear regression resulted in p-values < 0.00006 (or Bonferroni-adjusted p-values $p_a < 0.001$), two stars (**) indicate that tests for interactions using linear regression resulted in p-values < 0.0006 (or Bonferroni-adjusted p-values $p_a < 0.01$) and one star (*) indicates that tests for interactions using linear regression resulted in p-values < 0.003 (Bonferroni-adjusted p-values $p_a < 0.05$). Note, we interpret the resulting significant levels with caution and focus in our analysis mostly on the the importance of the effects to the model. We always provide contour plots of the actual data distributions of specific emerging interactions for each ASR architecture in order to analyze these more informative visualizations together with the explicit naming of a quantitative or qualitative interaction type and the corresponding significance levels.

In the remainder of this study, we will refer to interactions with significance level (***) as *highly significant* and interactions with significance level (**) or (*) as *significant*.

4.2.7.1 Statistical analysis of short utterances:

Given the WERs on utterance level of each ASR system of *short* utterances, we compared all EIM values of all six univariable effects and only the 3-best EIM values of the quantitative and qualitative interaction effects. Tab. 4.4 summarizes all EIM values with respect to the effect type (rows: univariable effects, quantitative interaction effects or qualitative interaction effects) and the ASR system (columns: Whisper, Kaldi, w2v and w2vLM). First, we describe the univariable EIM values (cf. Fig. 4.8) and after that we focus more precisely on the 3-best quantitative and the 3-best qualitative interaction effects. Consequently, for a better explanation of the most important interactions, we compare contour plots which collect mean WERs on utterance level with respect to specific grid areas and interpolate between values

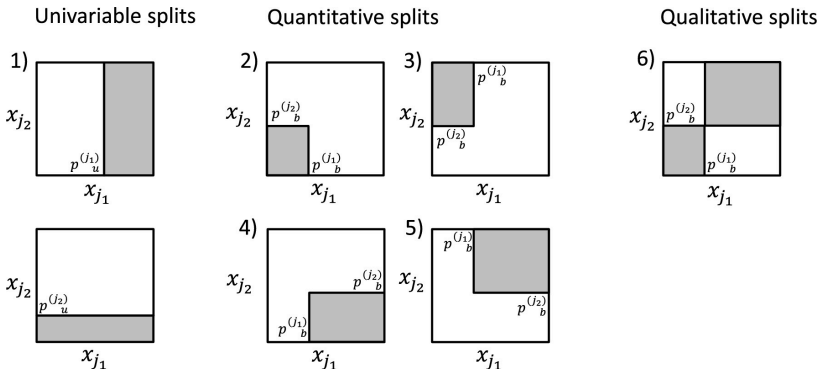


Figure 4.7: Split types considered in the Interaction Forest algorithm, figure taken from the original paper (Hornung & Boulesteix, 2022a). Split types comprise one univariable split, four quantitative splits and one qualitative split. Note that we describe our variables more generally as x and y without considering a particular variable j. Additionally, we describe split points p_u or p_b as x_s (split point for variable x) or y_s (split point for variable y).

with respect to equivalent color bars in order to give an easy and direct comparison between the performances of the ASR systems (to be more precise, this means that for a specific interaction the WERs on utterance level in the contour plots were all represented with the same color coding). The necessary grid areas were motivated by the two-dimensional LOESS fits produced with the R package `diversityForest` by capturing specific feature value ranges which where also conditioned on the feature type. More specifically, in case of the discrete feature **#tokens** we consider broader areas surrounding the integer values which enables a smooth interpolation between the mean WERs on utterance level (e.g., for utterances containing four word tokens the grid area was specified by a offset of 0.5 capturing the ranges $[0.5, 1.5)$, $[1.5, 2.5)$, $[2.5, 3.5)$ and $[3.5, 4.5)$ which allows a better readability by preventing blank areas in the contours plots) and in case of all other continuous features, we consider ranges which were linearly spaced with 10 steps between a minimum feature value specified as the 5%-quantile and a maximum feature value specified as the 95%-quantile (in this case the offsets for the grid areas were half of the difference between two resulting subsequent feature values). Given these grid areas and corresponding mean WERs on utterance level, all contour plots were automatically generated with the functions `contourf` and `contour` (only the parameter `levels` was adjusted to achieve comparable color bars while all other parameters were kept with the default settings) given the *matplotlib* package (version 3.6.2) developed for Python. As a results, we generated five sets of four contour plots for each ASR system (cf. Fig. 4.10, 4.11, 4.12, 4.13 and 4.14) when evaluating the 3-best quantitative interaction effects (cf. Tab. 4.4). Additionally, in case of the 3-best qualitative interaction effects (cf. Tab. 4.4) we generated another four sets of four contour plots (cf. Fig. 4.15, 4.16, 4.17 and 4.18).

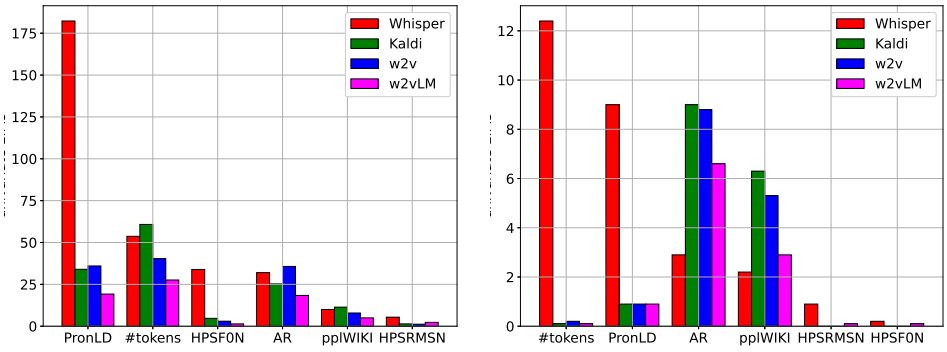
Univariable effects for *short* utterances: Tab. 4.4 and Fig. 4.8a summarize EIM values of the univariable effects of *short* utterances with respect to each ASR architecture. In case of Whisper, the feature **PronLD** was the most important feature with an EIM value of 182.3, followed by **#tokens** with an EIM value of 53.7.

Table 4.4: Summary of all univariable, 3-best quantitative and 3-best qualitative EIM values from the Interaction Forests (Hornung & Boulesteix, 2022a) for *short* utterances. Note that in this table the descriptions "Univ. Effects", "Quant. Inter." and "Qual. Inter." refer to univariable effects and the effects of quantitative and qualitative interactions. Additionally, (small) is abbreviated with (↓) and (large) with (↑).

	Whisper		**Kaldi**		**w2v**		**w2vLM**	
	Feature	**EIM**	**Feature**	**EIM**	**Feature**	**EIM**	**Feature**	**EIM**
Univ. Effects	**PronLD**	182.3	**#tokens**	60.8	**#tokens**	40.3	**#tokens**	27.6
	#tokens	53.7	**PronLD**	34	**PronLD**	36	**PronLD**	19.1
	HPSF0N	33.9	**AR**	25.3	**AR**	35.7	**AR**	18.4
	AR	32	**pplWIKI**	11.4	**pplWIKI**	7.9	**pplWIKI**	5
	pplWIKI	10	**HPSF0N**	4.7	**HPSF0N**	3	**HPSRMSN**$_{Res}$	2.3
	HPSRMSN$_{Res}$	5.4	**HPSRMSN**$_{Res}$	1.4	**HPSRMSN**$_{Res}$	1.2	**HPSF0N**	1.4
Quant. Inter.	**#tokens** (↓) **PronLD** (↓)	33.7***	**#tokens** (↓) **PronLD** (↓)	23.2***	**#tokens** (↓) **PronLD** (↓)	15.5***	**#tokens** (↓) **PronLD** (↓)	10.2***
	#tokens (↓) **HPSF0N** (↑)	26***	**#tokens** (↓) **AR** (↓)	12.9***	**#tokens** (↓) **AR** (↓)	7.7*	**#tokens** (↓) **pplWIKI** (↓)	5.1
	PronLD (↓) **HPSF0N** (↓)	14.3***	**#tokens** (↓) **pplWIKI** (↓)	10.2	**#tokens** (↓) **pplWIKI** (↓)	7.5*	**#tokens** (↓) **AR** (↓)	5
Qual. Inter.	**#tokens** **PronLD**	0.8***	**#tokens** **PronLD**	1.5***	**#tokens** **PronLD**	1.2***	**#tokens** **PronLD**	1.4***
	#tokens **HPSRMSN**$_{Res}$	0.5*	**PronLD** **AR**	0.2	**PronLD** **AR**	0.5	**PronLD** **AR**	0.3**
	HPSF0N **AR**	0.4***	**PronLD** **pplWIKI**	0.2*	**PronLD** **pplWIKI**	0.3**	**PronLD** **pplWIKI**	0.2*

In contrast, EIM values were ranked in a different order in case of Kaldi and the wav2vec2 architecture. Here, in all cases the most important feature was **#tokens** followed by the features **PronLD** and **AR**. The absolute differences between EIM values of the 3-best features were higher in case of Kaldi (28.8 and 8.7) in comparison to w2v (4.3 and 0.3) and w2vLM (8.5 and 0.7). To summarize, we recognize that the features **PronLD** and **#tokens** were the most important features across all ASR architectures for *short* utterances. However, in case of Whisper the feature **PronLD** had a much higher effect size in comparison to the second best feature **#tokens**. Additionally, the feature **AR** was also more important across all ASR architectures but the feature **HPSF0N** appeared to play a more important role only in the case of Whisper. In contrast, across all ASR architectures the remaining features **pplWIKI** and **HPSRMSN**$_{Res}$ seem be less important than the other features.

Fig. 4.9 provides a visualization of the univariable effects on the WERs on utterance level, illustrating the relationships between binned WERs on utterance level (with WER binning regions $[0\%, 10\%]$, $(10\%, 20\%]$, $(20\%, 30\%]$, $(30\%, 40\%]$, $(40\%, 50\%]$, $(50\%, 60\%]$, $(60\%, 70\%]$, $(70\%, 80\%]$, $(80\%, 90\%]$ and $(90\%, 100\%]$) and the mean feature values (rows: **AR**, **HPSF0N**, **HPSRMSN**$_{Res}$, **PronLD** and **pplWIKI**) with respect to the number of utterances achieving this WER (specified by the circle sizes), the number of tokens per utterance (red: one word token, orange: two word tokens, yellow: three word tokens and green: four word tokens) and each ASR architecture (columns: Whisper, Kaldi, w2v and w2vLM). For *single-word* utterances there are only two WERs (0% and 100%) resolvable. Naturally, WER resolutions get finer for two word tokens (0%, 50% and 100%), three word tokens (0%, 33.33%, 66.66% and 100%) or four word tokens (0%, 25%, 50%, 75% and 100%). Due to the fact that the distribution of WERs with respect to the number of utterances (specified by circle sizes) could be difficult to read, we refer back to



(a) Univariable EIMs for *short* utterances

(b) Univariable EIMs for *long* utterances

Figure 4.8: Bar plots of the univariable EIMs for (a) *short* utterances or (b) $5 \ldots 15$ word tokens for the 3 ASR architectures Whisper, Kaldi and wav2vec2 with and without lexicon/LM (w2v and w2vLM). Features were ordered with respect to the rankings of the univariable EIMs of the Whisper architecture. Note that the description of the feature **HPSRMSN** actually refers to the feature $\mathbf{HPSRMSN}_{Res}$ which was derived from residuals.

Fig. 4.4c and 4.4d in Sec. 4.2.6.1 which, i.a., showed that *single-word* utterances led to approx. 6000 (Whisper) and 8000 (Kaldi and the wav2vec2 architecture) WERs at 0% in contrast to approx. 4000 (Whisper) and 2000 (Kaldi and the wav2vec2 architecture) WERs at 100%. Therefore, in the following descriptions, we focus mainly on the mean feature values for specific WER binning regions instead of focusing on the WER distributions with respect to the number of utterances (specified by the circle sizes). One advantage of the visualization in Fig. 4.9 is that all features were displayed with respect to the utterance length feature **#tokens** which was an important feature across all ASR architectures. Thus, we explain clear trends of all univariable effects mainly with respect to the specific numbers of word tokens (i.e., one, two, three or four word tokens) in an utterance.

In case of the (more important) durational feature **AR** we predominantly observe positive trends irrespective of the number of word tokens per utterance indicating that higher articulation rates led to worse WERs. Especially in case of *single-word* utterances, we observe that WERs of 0% were achieved for mean articulation rates of $\approx 9\,\mathrm{s}^{-1}$ across all ASR architectures. In contrast, WERs of 100% were achieved for mean articulation rates of $\approx 10.5\,\mathrm{s}^{-1}$ (Whisper), $\approx 11\,\mathrm{s}^{-1}$ (Kaldi) and $\approx 11.5\,\mathrm{s}^{-1}$ (wav2vec2 architecture). In case of utterances with 2/3/4 word tokens the trends were similar but WERs of 0% were achieved for higher mean articulation rates of approx. $11\,\mathrm{s}^{-1}/12.5\,\mathrm{s}^{-1}/13\,\mathrm{s}^{-1}$ across all ASR architectures and WERs of 100% were achieved for even higher mean articulation rates of approx. $12\,\mathrm{s}^{-1}/13.5\,\mathrm{s}^{-1}/15\,\mathrm{s}^{-1}$ (Whisper), $12.5\,\mathrm{s}^{-1}/13.5\,\mathrm{s}^{-1}/14\,\mathrm{s}^{-1}$ (Kaldi), $12.5\,\mathrm{s}^{-1}/14\,\mathrm{s}^{-1}/14.5\,\mathrm{s}^{-1}$ (w2v) and $12.5\,\mathrm{s}^{-1}/14\,\mathrm{s}^{-1}/15\,\mathrm{s}^{-1}$ (w2vLM). Nevertheless, in case of Whisper and Kaldi, WERs of 50% were achieved for lower mean articulation rates of approx. $10.5\,\mathrm{s}^{-1}/11\,\mathrm{s}^{-1}$ in contrast to the wav2vec2 architecture which achieved a mean articulation rate of $\approx 11.5\,\mathrm{s}^{-1}$ in both cases (w2v and w2vLM).

In case of the pseudo-entropy of the F0 contour **HPSF0N** (which was more important only in case of Whisper), we observe the opposite trend indicating that more uniformly distributed F0 contours led to better WERs. Indeed, for *single-*

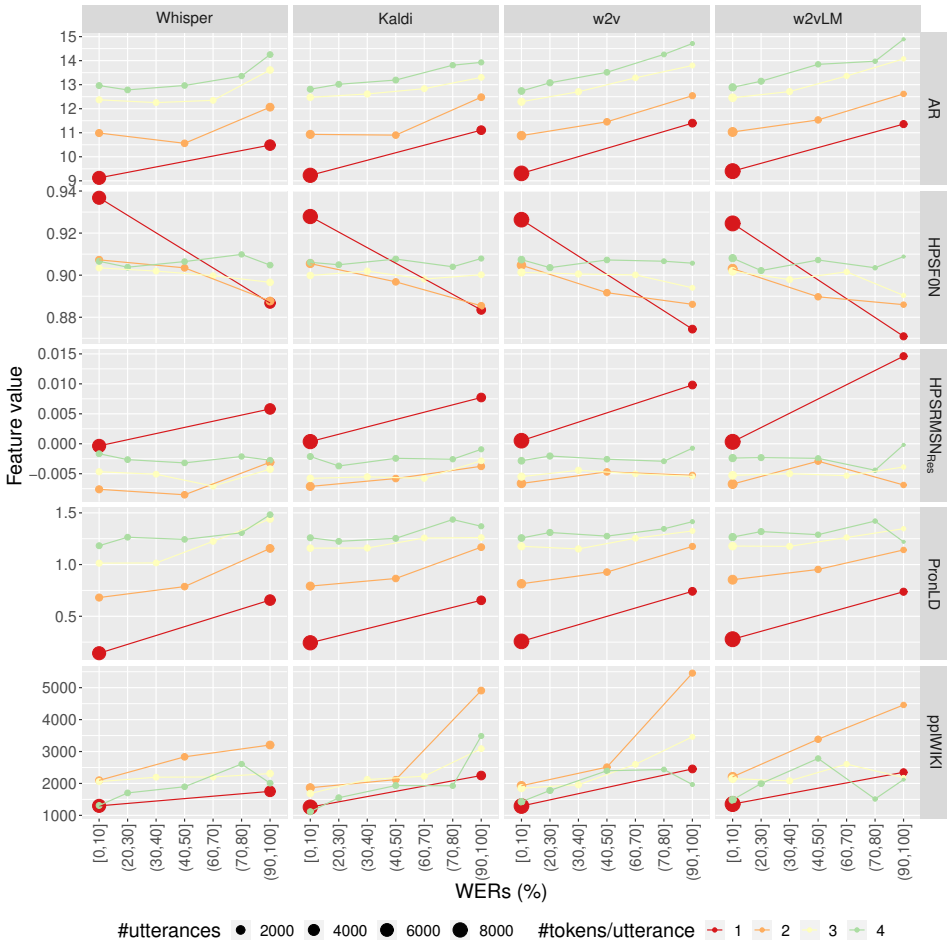


Figure 4.9: Relationships between WERs [%] on utterance level and mean feature values of selected higher correlating features (rows: **AR**, **HPSF0N**, **HPSRMSN**$_{Res}$, **PronLD** and **pplWIKI**) with respect to each ASR system (columns: Whisper, Kaldi, w2v and w2vLM). WERs of 10%-intervals are summarized and conditioned on the number of utterances (depicted by circles sizes) as well as the number of tokens per utterance (i.e., colors red (one word token), orange (two word tokens), yellow (three word tokens) and green (four word tokens)). For utterances containing one word token there are only two WERs (0% and 100%) possible. WER resolutions get finer for two word tokens (0%, 50% and 100%), three word tokens (0%, 33.33%, 66.66% and 100%) or four word tokens (0%, 25%, 50%, 75% and 100%).

word utterances, WERs of 0% were achieved for higher mean pseudo-entropies of approx. 0.94/0.93/0.93/0.92 (Whisper/Kaldi/w2v/w2vLM) and WERs of 100% were achieved for lower mean pseudo-entropies of approx. 0.89/0.88/0.87/0.87 (Whisper/Kaldi/w2v/w2vLM). Hence, this trend indicates better WERs for more uniformly distributed F0 contours (which corresponds to higher values of **HPSF0N**) in case of *single-word* utterances. With respect to the utterances with two word tokens, better WERs across all ASR architectures tended to occur together with higher mean pseudo-entropies. For utterances with $3-4$ word tokens, there is no clear trend for whether **HPSF0N** affects the performance

of any of the ASR architectures.

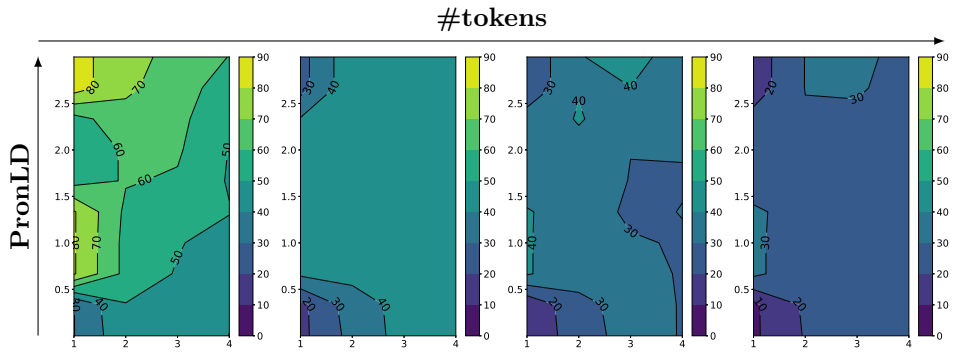
With respect to the RMS Feature, the residuals $\mathbf{HPSRMSN}_{Res}$ from a linear model (cf. equations (4.16) and (4.17)) showed a clear positive trend merely in case of *single-word* utterances. In this case, WERs of 0% were achieved if $\mathbf{HPSRMSN}_{Res}$ a mean value of ≈ 0 across all ASR architectures and WERs of 100% were achieved for slighlty higher mean values of 0.005/0.0075/0.01/0.015 (Whisper/Kaldi/w2v/w2vLM). Note that, in contrast to **HPSF0N**, this feature is not easy to interpret intuitively, as a large (small) value of the residual of the linear model does not necessarily imply that the original feature **HPSRMSN** was large (small), or that the RMS contour was flat (varying). For utterances with $3-4$ word tokens all values of $\mathbf{HPSRMSN}_{Res}$ were < 0 independent of the ASR architecture and we observe no clear trends.

In case of the (more important) pronunciation feature **PronLD** (best univariable effect in case of Whisper and second best in case of Kaldi and the wav2vec2 architecture) which, simply put, describes the degree to which specific words were spoken dialectically (in the sense of pronunciation reduction), we observe clear trends independent of the number of word tokens in an utterance. In case of *single-word* utterances mean values of **PronLD** of approx. 0 (Whisper) or 0.25 (Kaldi and the wav2vec2 architecture) mainly resulted in WERs of 0% in contrast to mean values of ≈ 0.75 which resulted in WERs of 100%. In case of utterances with 2/3/4 word tokens the trends were also positive but WERs of 0% were achieved for higher mean values of **PronLD** of approx. $0.75/1-1.25/1.25$ across all ASR architectures and WERs of 100% were achieved for even higher mean values of **PronLD** of approx. 1.25/1.5/1.5 (Whisper), 1.25/1.25/1.25 (Kaldi), 1.25/1.5/1.25 (w2v) and 1.25/1.25/1.25 (w2vLM). Utterances with four word tokens demonstrated a higher mean value for **PronLD** of ≈ 1.5 (w2v) in comparison to ≈ 1.25 (w2vLM) but then again in case of w2vLM there were also fewer utterances affected (as illustrated via the smaller circle size).

Finally, for **pplWIKI** there were stronger differences between the ASR architectures especially for utterances containing two word tokens, but in general these trends were positive (indicating worse WERs of higher perplexities) but also weaker than for other features. Most salient, perhaps, is the behavior for utterances with two word tokens, where the positive trend of Whisper is much flatter than for other ASR architectures, and where the wav2vec2 architecture without a lexicon/LM achieved WERs of 100% primarily for utterances that had very high perplexities (≈ 5000 versus ≈ 4500).

Quantitative interaction effects for *short* utterances: Tab. 4.4 summarizes the 3-best quantitative interactions across all ASR architectures. First, we summarize the corresponding EIM values of these quantitative interactions and after that, we analyze all interaction effects individually by describing corresponding contour plots. The quantitative (and also qualitative) interaction between **#tokens** (small) and **PronLD** (small)[5] was the most important interaction effect across all ASR architectures with EIM values of 33.7 (Whisper), 23.2 (Kaldi), 15.5 (w2v) and 10.2

[5]Here and subsequently, the additions (small) and (large) indicate where strongest interactions take place. For example, **feature 1** (small) and **feature 2** (large) indicates that the WERs change more strongly with **feature 2** if **feature 1** is small, or more strongly with **feature 1** if **feature 2** is large.

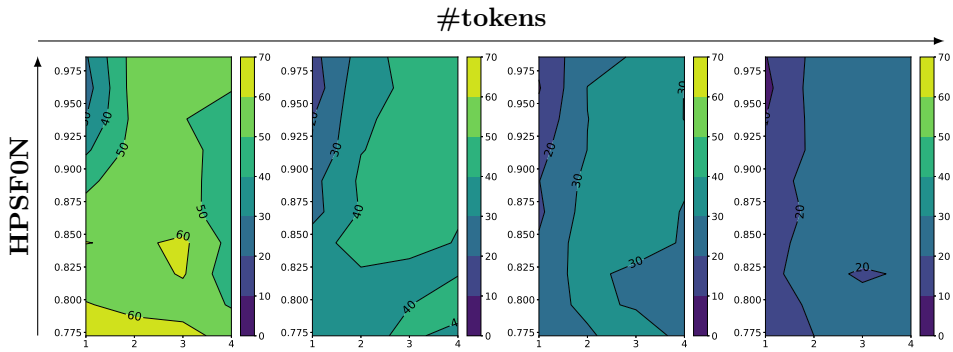


(a) Whisper*** (b) Kaldi*** (c) w2v*** (d) w2vLM***

Figure 4.10: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...90% reflects mean WERs on utterance level within a grid area. Visualizations refer to the quantitative (and qualitative) interaction between **#tokens** (small) and **PronLD** (small) across all ASR architectures. All tests for interaction using linear regression had significant p-values < 0.00006 (***).

(w2vLM). Additionally, all interactions were *highly significant*. Interestingly, this first and best quantitative interaction was also the most important qualitative interaction for all ASR systems, albeit with a much smaller effect size. With respect to the EIM value, the quantitative interaction between **#tokens** (small) and **HPSF0N** (large) was more important in case of Whisper with an EIM value of 26. Apart from that, tests for interactions were *highly significant* for Whisper, Kaldi and w2v and *significant* for w2vLM. The quantitative interaction between **PronLD** (small) and **HPSF0N** (large) resulted to affect WERs of Whisper with an EIM value of 14.3. Interactions in case of Whisper, Kaldi and w2v were *highly significant* and for w2vLM the interaction was not significant. The quantitative interaction between **#tokens** (small) and **AR** (small) showed to significantly affect the performance of Kaldi (EIM value of 12.9), but less so for w2v and w2vLM (EIM values of 7.7 and 5.1). Here, interactions in case of Whisper and Kaldi were *highly significant* and in case of the wav2vec2 architecture the interactions were not significant. The quantitative interaction between **#tokens** (small) and **pplWIKI** (small) tended to be stronger for Kaldi (EIM value of 10.2) than for w2v and w2vLM (EIM values of 7.5 and 5). Nevertheless, only the interaction for w2v turned out to be *significant*.

For all ASR architectures, the performance was most strongly affected by the quantitative interaction between **#tokens** (small) and **PronLD** (small), which was *highly significant*. Fig. 4.10 illustrates that a small number of **#tokens** (e.g., one word token) and a small number of **PronLD** (closer to the canonical pronunciation) affect the WERs. Simultaneously, the WER tends to change strongly with **PronLD** if **#tokens** is small and only weakly if **#tokens** is large. Generally, we observe that all architectures achieved best mean WERs of approx. $20\% - 30\%/10\% - 20\%/10\% - 20\%/0\% - 10\%$ (Whisper/Kaldi/w2v/w2vLM) for *single-word* utterances and a **PronLD** of 0 (note that in case of one word token the feature **PronLD** has only integer values). Kaldi and w2v achieved similar mean WERs for lower and higher values of **PronLD** for one word token ($10\% - 20\%$ and $20\% - 30\%$). In contrast to all other architectures, Whisper achieved high mean WERs of approx. $80\% - 90\%$



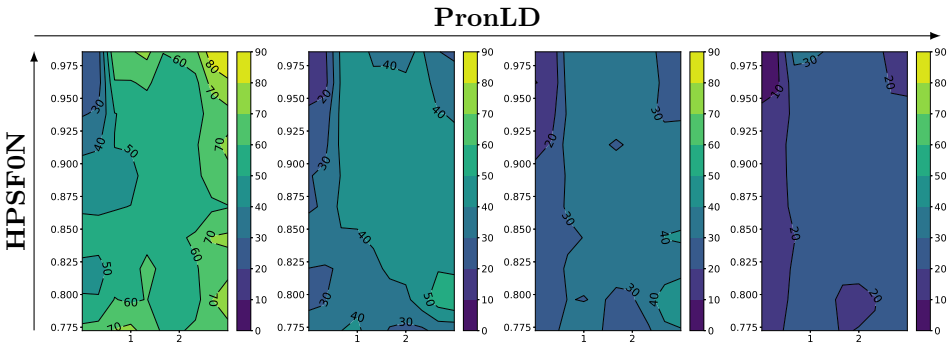
(a) Whisper*** (b) Kaldi*** (c) w2v*** (d) w2vLM*

Figure 4.11: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...60% reflects mean WERs on utterance level within a grid area. Visualizations refer to the quantitative interaction between **#tokens** (small) and **HPSF0N** (large) across all ASR architectures. All tests for interaction resulted in significant p-values. Tests for interaction in case of Whisper, Kaldi and w2v using linear regression were significant with p-values < 0.00006 (***). In case of w2vLM the interaction was also significant with a p-value < 0.003 (*).

for *single-word* utterances and higher values of **PronLD**. Kaldi and w2vLM were more robust and almost independent of **PronLD** for utterances containing only two to four word tokens (mean WERs were between $40\% - 50\%$ and $20\% - 30\%$). Finally, both wav2vec2 systems had worse WERs for two to three word tokens and generally higher values of **PronLD** (approx. $40\% - 50\%$ (w2v) and $30\% - 40\%$ (w2vLM))

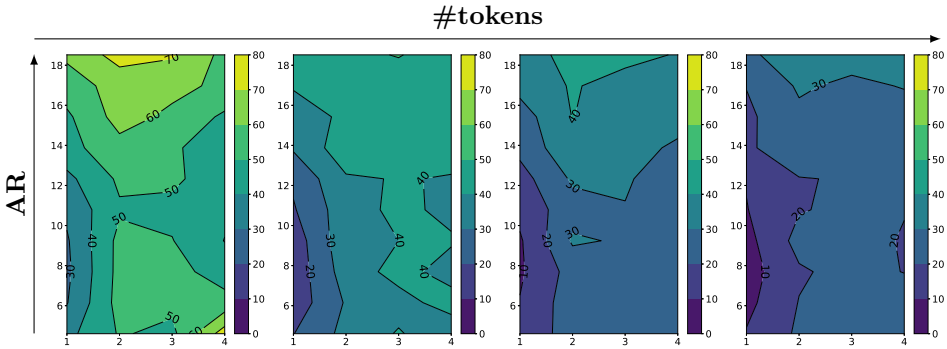
The second strongest quantitative interaction effect on the utterance-level WERs was between **#tokens** (small) and **HPSF0N** (large), which resulted to be *highly significant* for Whisper, Kaldi and w2v and *significant* for w2vLM. Fig. 4.11 illustrates that for Whisper, the mean WERs for one word token were between $10\% - 20\%$ for high values of **HPSF0N** (more uniformly distributed F0 contours) and between $60\% - 70\%$ for lower values (less uniformly distributed F0 contours). For Kaldi and w2v, for the same comparison, the differences were between $10\% - 20\%$ and $30\% - 40\%$ and $10\% - 20\%$ and $20\% - 30\%$ respectively. For two to three word tokens Whisper achieved mean WERs between $50\% - 70\%$ independent of **HPSF0N** and for four word tokens between $40\% - 60\%$. For all other architectures, with a few exceptions, these WERs for two to four word tokens tended to be roughly between $40\% - 50\%$ (Kaldi), $30\% - 40\%$ (w2v) and $20\% - 30\%$ (w2vLM), with the distribution of WERs being particularly uniform for w2vLM.

Fig. 4.12 shows the third quantitative interaction between **PronLD** (small) and **HPSF0N** (large) which resulted to have a *highly significant* effect for Whisper, Kaldi and w2v. This quantitative interaction was particularly visible in the case of Whisper, with best WERs between $20\% - 30\%$ for small values of **PronLD** (closer to the canonical pronunciation) and large values of **HPSF0N** (more uniformly distributed F0 contours), but only worse WERs between $60\% - 90\%$ for higher values of **PronLD** (further away from the canonical pronunciation). For all other architectures, the quantitative interaction related to the specified range also applied with WERs between $10\% - 20\%$ (Kaldi/w2v) and $0\% - 10\%$ (w2vLM). In contrast, it



(a) Whisper*** (b) Kaldi*** (c) w2v*** (d) w2vLM

Figure 4.12: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...90% reflects mean WERs on utterance level within a grid area. Visualizations refer to the quantitative interaction between **PronLD** (small) and **HPSF0N** (large) across all ASR architectures. Tests for interaction in case of Whisper, Kaldi and w2v using linear regression were significant with p-values < 0.00006 (***). In case of w2vLM the interaction was not significant with a p-value $\geqslant 0.003$.

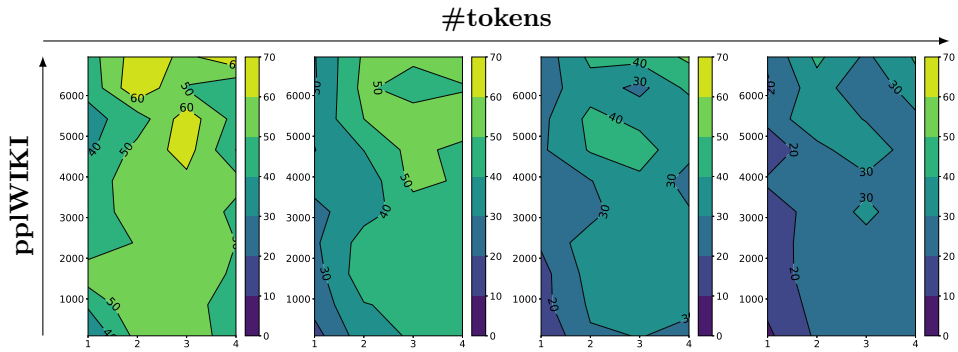


(a) Whisper*** (b) Kaldi*** (c) w2v (d) w2vLM

Figure 4.13: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...80% reflects mean WERs on utterance level within a grid area. Visualizations refer to the quantitative interaction between **#tokens** (small) and **AR** (small) across all ASR architectures. Tests for interaction in case of Whisper and Kaldi using linear regression were significant with p-values < 0.00006 (***). In case of the wav2vec2 architecture the interactions were not significant with p-values $\geqslant 0.003$.

can be stated that for decreasing **HPSF0N** (less uniformly distributed F0 contours) and increasing **PronLD** (further away from the canonical pronunciation) generally the WERs were between $40\% - 90\%$ (Whisper), $20\% - 60\%$ (Kaldi), $20\% - 50\%$ (w2v) and $10\% - 40\%$ (w2vLM).

Fig. 4.13 illustrates the fourth quantitative interaction between **#tokens** (small) and **AR** (small), which was *highly significant* for Whisper and Kaldi. In general, WERs were best for articulation rates $\leqslant 10\,\mathrm{s}^{-1}$ and one word token (approx. $20\% - 30\%$ (Whisper), $10\% - 20\%$ (Kaldi/w2v) and $0\% - 10\%$ (w2vLM)). However,



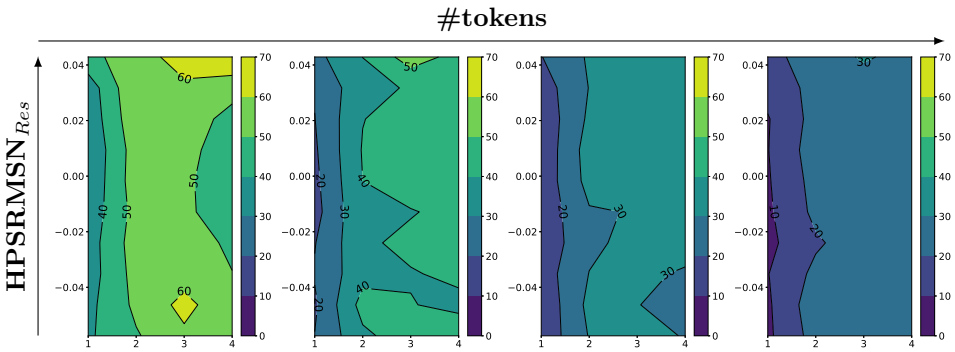
(a) Whisper (b) Kaldi (c) w2v* (d) w2vLM

Figure 4.14: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...70% reflects mean WERs on utterance level within a grid area. Visualizations refer to the quantitative interaction between **#tokens** (small) and **pplWIKI** (small) across all ASR architectures. Tests for interaction in case of w2v using linear regression were significant with a p-value < 0.003 (*). In case of Kaldi, Whisper and w2vLM the interactions were not significant with p-values $\geqslant 0.003$.

for two and three word tokens, Whisper showed that a kind of *sweet spot* emerged for articulation rates between approx. $10\,\mathrm{s}^{-1} - 1212\,\mathrm{s}^{-1}$. Regardless of the number of words, all architectures showed higher WERs for higher articulation rates of approx. $\geqslant 12\,\mathrm{s}^{-1}$ (approx. $40\% - 80\%$ (Whisper), $20\% - 50\%$ (Kaldi), $10\% - 50\%$ (w2v) and $10\% - 40\%$ (w2vLM)). Kaldi achieved WERs of approx. $30\% - 40\%$ for two word tokens and lower articulation rates ($< 12\,\mathrm{s}^{-1}$), but worse WERs of approx. $30\% - 50\%$ for three and four word tokens. For wav2vec2, the WERs for two to four word tokens were more uniformly distributed and the integration of a lexicon/LM led to a constant performance improvement.

Fig. 4.14 illustrates the fifth and final quantitative interaction between **#tokens** (small) and **pplWIKI** (small), which resulted to be *significant* only for w2v. In general, WERs were worse for higher perplexities (approx. **pplWIKI** $\geqslant 4000$) with approx. $30\% - 70\%$ (Whisper), $30\% - 60\%$ (Kaldi), $20\% - 50\%$ (w2v) and $10\% - 40\%$ (w2vLM). In contrast to the other systems, Whisper's performance was better for utterances of four word tokens than for those with two and three word tokens.

Qualitative interaction effects for *short* utterances: Tab. 4.4 summarizes the 3-best qualitative interactions across all ASR architectures. First, we summarize the corresponding EIM values of these qualitative interactions and after that, we analyze all interaction effects individually by describing corresponding contour plots. As already mentioned the best quantitative interaction between **#tokens** (small) and **PronLD** (small) was also the best qualitative interaction effect with EIM values of 0.8 (Whisper), 1.5 (Kaldi), 1.2 (w2v) and 1.4 (w2vLM). All other EIM values indicated less important interactions with values $\leqslant 0.5$. The qualitative interaction between **#tokens** and $\mathbf{HPSRMSN}_{Res}$ showed an interaction effect only in case of Whisper with an EIM value of 0.5. These interactions were *highly significant* for Whisper and w2v and *significant* for w2vLM. In case of Kaldi this interaction was not significant. Likewise, the qualitative interaction between **HPSF0N** and



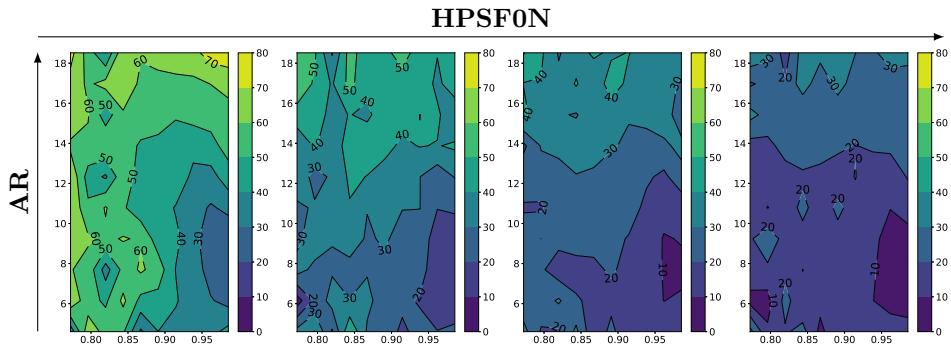
(a) Whisper* (b) Kaldi (c) w2v* (d) w2vLM***

Figure 4.15: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...70% reflects mean WERs on utterance level within a grid area. Visualizations refer to the qualitative interaction between **#tokens** and **HPSRMSN**$_{Res}$ across all ASR architectures. Tests for interaction in case of Whisper and the wav2vec2 architecture using linear regression were significant with p-values < 0.00006 (***) in case of Whisper/w2vLM and a p-value < 0.003 (*) in case of w2v. In case of Kaldi the interaction was not significant with a p-value $\geqslant 0.003$.

AR demonstrated an interaction effect only for Whisper with an EIM value of 0.4. In this case, interactions were *highly significant* for all ASR architectures. The qualitative interaction between **PronLD** and **AR** showed interaction effects in case of the trained or fine-tuned architectures with EIM values of 0.2 (Kaldi), 0.5 (w2v) and 0.3 (w2vLM). The interaction in case of Whisper was *highly significant* and for w2vLM *significant* but for Kaldi and w2v the interactions were not significant. The final qualitative interaction between **PronLD** and **pplWIKI** showed again interaction effects in case of the trained or fine-tuned architectures with EIM values of 0.2 (Kaldi), 0.3 (w2v) and 0.2 (w2vLM). In this case, interactions were *significant* across all ASR architectures.

Fig. 4.15 illustrates the first qualitative interaction between **#tokens** and **HPSRMSN**$_{Res}$ which was *highly significant* for w2vLM and *significant* for Whisper and w2v. This qualitative interaction effect indicates that the WERs increase with the number of word tokens for higher values of **HPSRMSN**$_{Res}$ but decrease for lower values of **HPSRMSN**$_{Res}$. We notice a slight effect of this with Whisper as the WERs for utterances containing one word token and higher values of **HPSRMSN**$_{Res}$ were approx. between $30\% - 40\%$ and for four word tokens approx. between $60\% - 70\%$. Similarly, for utterances containing three word tokens and some lower values of **HPSRMSN**$_{Res}$ at ≈ -0.05 the WERs were approx. between $60\% - 70\%$ and then decreased for four word tokens to $50\% - 60\%$. Nonetheless, this effect was rather minimal for Whisper and hardly present in case of the other architectures. Then again, across all ASR architectures, we generally observe that the WERs got worse with an increasing number of word tokens independent of **HPSRMSN**$_{Res}$. The only exception was Whisper, where better WERs were achieved again with four word tokens and average values of **HPSRMSN**$_{Res}$ between approx. $-0.02 \ldots 0.02$.

Fig. 4.16 illustrates the second qualitative interaction between **HPSF0N** and **AR** which was *highly significant* in case of all ASR architectures. With respect to



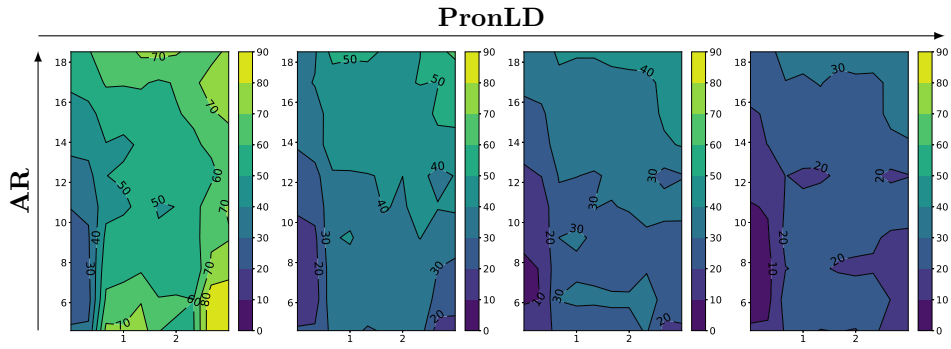
(a) Whisper*** (b) Kaldi*** (c) w2v*** (d) w2vLM***

Figure 4.16: Contour plots of the actual data distributions of *short* utterances. The color map between 0% ... 80% reflects mean WERs on utterance level within a grid area. Visualizations refer to the qualitative interaction between **HPSF0N** and **AR** across all ASR architectures. All tests for interaction using linear regression were significant with p-values < 0.00006 (***).

the contour plots, in this case the qualitative interaction was not really visible in any of the ASR systems. However, across all ASR architectures for higher values of **HPSF0N** (more uniformly distributed F0 contours) and lower articulation rates (between approx. $6\,\mathrm{s}^{-1} \ldots 10\,\mathrm{s}^{-1}$) better WERs of $20\% - 30\%$ (Whisper), $10\% - 20\%$ (Kaldi) and $0\% - 10\%$ (wav2vec2) were achieved (with w2v being a small exception with articulation rates of $\approx 8\,\mathrm{s}^{-1}$). Generally, in case of Kaldi and the wav2vec2 architecture the WERs got worse with increasing articulation rates and this rather independently of the distribution of the F0 contour. In principle, Whisper also showed that worse WERs (approx. between $60\% - 70\%$) were achieved for lower values of **HPSF0N** and this almost independently of the articulation rates. At the same time, Whisper achieved worse results for large articulation rates and large values of **HPSF0N** (between $70\% - 80\%$).

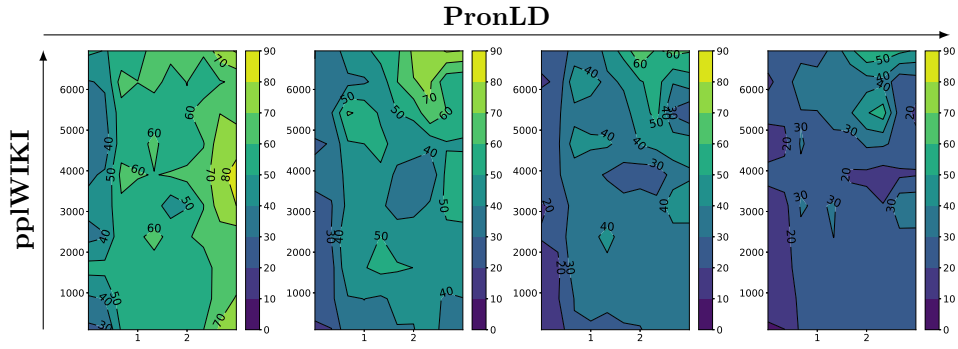
Fig. 4.17 illustrates the third qualitative interaction between **PronLD** and **AR** which was *highly significant* for Whisper and *significant* for w2vLM. This qualitative interaction effect indicates that the WERs tend to increase with the articulation rate for lower values of **PronLD** (closer to the canonical pronunciation) but tend to decrease with a large number of **PronLD** (further away from the canonical pronunciation). This qualitative interaction was more or less observable only in case of Whisper. Across all ASR architectures, WERs were best for articulation rates between approx. $5\,\mathrm{s}^{-1} \ldots 10\,\mathrm{s}^{-1}$ and **PronLD** close to 0 (with w2v being again a small exception with articulation rates of approx. $6\,\mathrm{s}^{-1} \ldots 8\,\mathrm{s}^{-1}$). For Kaldi and the wav2vec2 architecture, better WERs between $10\% - 20\%$ were also achieved for small articulation rates (approx. $< 5\,\mathrm{s}^{-1}$) which were further away from the canonical pronunciation (**PronLD** ≈ 3). Furthermore, for Kaldi and the wav2vec2 architecture, WERs between $40\% - 50\%$ (Kaldi), $30\% - 50\%$ (w2vLM) and $20\% - 40\%$ (w2v) were achieved for articulation rates of approx. $> 12\,\mathrm{s}^{-1}$ and **PronLD** of approx. > 0.5.

Fig. 4.18 illustrates the fourth and last qualitative interaction between **PronLD** and **pplWIKI** which was *highly significant* for Whisper and w2v and *significant* for Kaldi and w2vLM. This qualitative interaction was not really visible in any of



(a) Whisper*** (b) Kaldi (c) w2v (d) w2vLM**

Figure 4.17: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...100% reflects mean WERs on utterance level within a grid area. Visualizations refer to the qualitative interaction between **PronLD** and **AR** across all ASR architectures. Tests for interaction in case of Whisper and w2vLM using linear regression were significant with a p-value < 0.00006 (***) in case of Whisper and a p-value < 0.0006 (**) in case of w2vLM. In case of Kaldi and w2v the interactions were not significant with p-values $\geqslant 0.003$.



(a) Whisper** (b) Kaldi* (c) w2v** (d) w2vLM*

Figure 4.18: Contour plots of the actual data distributions of *short* utterances. The color map between 0%...90% reflects mean WERs on utterance level within a grid area. Visualizations refer to the qualitative interaction between **PronLD** and **pplWIKI** across all ASR architectures. All tests for interaction using linear regression were significant with p-values < 0.0006 (**) in case of Whisper and w2v and p-values < 0.003 (*) in case of Kaldi and w2vLM.

the ASR systems but at least for Kaldi and the wav2vec2 architecture, we observed general trends for lower values of **PronLD** (closer to the canonical pronunciation) and lower perplexities as well as higher values of **PronLD** (further away from the canonical pronunciation) and higher perplexities. In particular, for lower values the WERs were better between $10\% - 30\%$ (Kaldi), $10\% - 20\%$ (wav2vec2) and for higher values the WERs were worse between $60\% - 70\%$ (Kaldi), $50\% - 60\%$ (w2v) and $40\% - 50\%$ (w2vLM). In case of Whisper the interaction was more complex but in general, we found that worse WERs between $60\% - 90\%$ were achieved for pronunciations further away from the canonical pronunciation.

4.2.7.2 Statistical analysis of long utterances:

Given the WERs on utterance level of each ASR system of *long* utterances, we compared all EIM values of all six univariable effects and only the 3-best EIM values of the quantitative and qualitative interaction effects. Tab. 4.5 summarizes all EIM values with respect to the effect type (rows: univariable effects, quantitative interaction effects or qualitative interaction effects) and the ASR system (columns: Whisper, Kaldi, w2v and w2vLM). In case of the *long* utterances our analysis focuses solely on the univariable EIM values (cf. Fig. 4.8). This is because both quantitative and qualitative interaction effects resulted in very low EIM values and we also found that corresponding contour plots confirmed less important interactions as general trends were difficult or even impossible to identify. In particular, it turned out that EIM values for quantitative and qualitative interaction effects were between $0.4 - 3$ and $0.02 - 0.11$. This prompted us to take a closer look only at the univariable effects, as it cannot be ruled out that we might not be able to make generalizable statements when describing the (rather unimportant) interactions.

Univariable effects for *long* utterances: Tab. 4.5 and Fig. 4.8b summarize EIM values of the univariable effects of *long* utterances with respect to each ASR architecture. In this case, Whisper ranked the feature **#tokens** as the most important feature with an EIM value of 12.4 (cf. Fig. 4.5) and the feature **PronLD** as the second best feature with an EIM value of 9. All other features were less important leading to EIM values of 2.9 (**AR**), 2.2 (**pplWIKI**) 0.9 ($\mathbf{HPSRMSN}_{Res}$) and 0.2 (**pplWIKI**). In contrast, best features in case of Kaldi and the wav2vec2 architecture were **AR** and **pplWIKI** which achieved EIM values of 9 and 6.3 (Kaldi), 8.8 and 5.3 (w2v) and 6.6 and 2.9 (w2vLM). All other remaining EIM values were similarly important with EIM values $\leqslant 0.9$. The absolute differences between EIM values of the 3-best features were higher for Kaldi (28.8 and 8.7) in comparison to w2v (4.3 and 0.3) and w2vLM (8.5 and 0.7). For Kaldi, w2v and w2vLM, the features **HPSF0N** and $\mathbf{HPSRMSN}_{Res}$ had EIM values close to 0.

Fig. 4.19 illustrates the univariable effects on the WERs on utterance level. This visualisation is similar to Fig. 4.9 and shows again the relationships between binned WERs on utterance level and the mean feature values (rows: **#tokens**, **AR**, **HPSF0N**, $\mathbf{HPSRMSN}_{Res}$, **PronLD** and **pplWIKI**) with respect to the number of utterances achieving this WER (specified by the circle sizes) and each ASR system (columns: Whisper, Kaldi, w2v and w2vLM). Note that here we also visualized the utterance length feature **#tokens** since we are no longer visualize the trends dependent on the number of word tokens, but rather summarizing them as averages.

In case of the utterance length feature **#tokens** which was more important for Whisper the WERs were best (between $0\% - 20\%$) for utterances containing approx. eight word tokens (Whisper/w2v/w2vLM) or approx. seven to eight word tokens (Kaldi). In contrast, worst WERs (between $70\% - 100\%$) were achieved for utterances containing approx. six to eight word tokens (Whisper/Kaldi/w2v) or approx. five to eight word tokens (w2vLM). However, in the case of the wav2vec2 architecture the worse WERs related to fewer utterances.

For the durational feature **AR**, which was more important for Kaldi and the wav2vec2 architecture, we found that in case of w2v and w2vLM the fewer utterances that were misrecognized had WERs between $40\% - 100\%$ for higher articulation

Table 4.5: Summary of all univariable, 3-best quantitative and 3-best qualitative EIM values from the Interaction Forests (Hornung & Boulesteix, 2022a) for *long* utterances. Note that in this table the descriptions "Univ. Effects", "Quant. Inter." and "Qual. Inter." refer to univariable effects and the effects of quantitative and qualitative interactions. Additionally, (small) is abbreviated with (↓) and (large) with (↑).

	Whisper		**Kaldi**		**w2v**		**w2vLM**	
	Feature	**EIM**	**Feature**	**EIM**	**Feature**	**EIM**	**Feature**	**EIM**
Univ. Effects	**#tokens**	12.4	**AR**	9	**AR**	8.8	**AR**	6.6
	PronLD	9	**pplWIKI**	6.3	**pplWIKI**	5.3	**pplWIKI**	2.9
	AR	2.9	**PronLD**	0.9	**PronLD**	0.9	**PronLD**	0.9
	pplWIKI	2.2	**#tokens**	0.1	**#tokens**	0.2	**HPSF0N**	0.1
	HPSRMSN$_{Res}$	0.9	**HPSF0N**	0	**HPSF0N**	0	**HPSRMSN**$_{Res}$	0.1
	HPSF0N	0.2	**HPSRMSN**$_{Res}$	0	**HPSRMSN**$_{Res}$	0	**#tokens**	0.1
Quant. Inter.	**#tokens** (↓) **PronLD** (↓)	3	**PronLD** (↓) **pplWIKI** (↓)	1	**PronLD** (↓) **pplWIKI** (↓)	1	**PronLD** (↓) **pplWIKI** (↓)	0.7
	PronLD (↓) **pplWIKI** (↓)	1.8	**AR** (↓) **pplWIKI** (↓)	1	**AR** (↓) **pplWIKI** (↓)	0.9**	**#tokens** (↓) **AR** (↓)	0.6
	PronLD (↓) **HPSRMSN**$_{Res}$ (↓)	0.8*	**#tokens** (↓) **AR** (↓)	0.7	**#tokens** (↓) **AR** (↓)	0.8	**AR** (↓) **pplWIKI** (↓)	0.4*
Qual. Inter.	**#tokens** **PronLD**	0.11	**PronLD** **AR**	0.03	**PronLD** **pplWIKI**	0.04	**AR** **pplWIKI**	0.03*
	PronLD **pplWIKI**	0.06	**PronLD** **pplWIKI**	0.03	**AR** **pplWIKI**	0.04**	**PronLD** **pplWIKI**	0.03
	#tokens **pplWIKI**	0.04	**AR** **pplWIKI**	0.03	**HPSRMSN**$_{Res}$ **pplWIKI**	0.03	**PronLD** **AR**	0.02

rates of approx. $14.5\,\mathrm{s}^{-1} \ldots 16\,\mathrm{s}^{-1}$. Conversely, for Whisper and Kaldi worse WERs (between $80\% - 100\%$) were achieved for a slightly higher number of utterances with articulation rates of approx. $14.5\,\mathrm{s}^{-1} \ldots 15\,\mathrm{s}^{-1}$. In general, across all ASR architectures best WERs (between $0\% - 40\%$) were achieved for lower articulation rates of approx. $13.5\,\mathrm{s}^{-1} \ldots 14.5\,\mathrm{s}^{-1}$.

In case of the (rather unimportant) prosodic features **HPSF0N** and **HPSRMSN**$_{Res}$ it was more difficult to observe indicative trends. For **HPSF0N**, there were at least minor anomalies in case of Kaldi (WERs between $70\% - 100\%$ for **HPSF0N** of approx. < 0.92) and w2v (WERs between $70\% - 100\%$ for **HPSF0N** of approx. < 0.918). In case of **HPSRMSN**$_{Res}$ there was also a slightly negative trend. Nevertheless, one should interpret these results with caution because the mean feature values did not really follow a clear trend. Overall, the prosodic features for *long* utterances showed little indication of important relationships with the WERs on utterance level.

The pronunciation feature **PronLD** was most important for Whisper which can be confirmed by the observation of a clear positive trend. Consequently, for Whisper the WERs were best between $0\% - 20\%$ for **PronLD** values of approx. < 1.35 (closer to canonical pronunciation) and worse between $80\% - 100\%$ for higher **PronLD** values of approx. > 1.6 (further away from the canonical pronunciation).

Finally for the (less important) perplexity feature **pplWIKI** we also observe slightly positive trends across all ASR architectures. This trend was weaker for Whisper where better WERs ($< 50\%$) were achieved for perplexities between approx. $1500 \ldots 1750$ and worse WERs ($\geqslant 50\%$) for perplexities between approx. $2000 \ldots 25000$. For Kaldi this trend was similar but also more linear. Conversely, for the wav2vec2 architecture worse WERs ($\geqslant 90\%$) were achieved for higher perplexities of approx. 3000 (yet again, overall fewer utterances were impacted

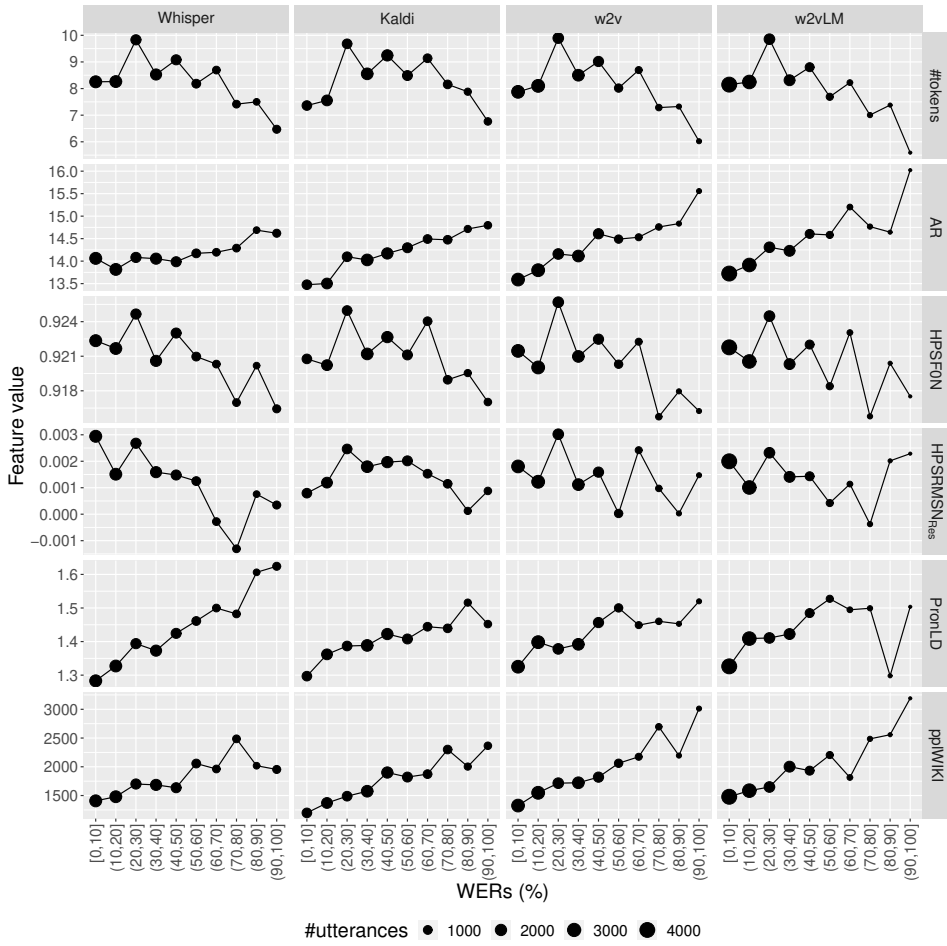


Figure 4.19: Relationships between feature values of selected higher correlating features (rows: **AR**, **#tokens**, **HPSF0N**, **HPSRMSN**$_{Res}$, **PronLD** and **pplWIKI**) and WERs [%] on utterance level with respect to each ASR system (columns: Whisper, Kaldi, w2v and w2vLM). WERs of 10%-intervals are summarized and conditioned on the number of utterances (depicted by circles sizes) but in this case independent of the number of word tokens.

by these mean values). Interestingly, for Kaldi and w2v best WERs $< 10\%$ were achieved at comparatively low mean perplexities of < 1500, for w2vLM, however, the corresponding mean perplexity value was higher (at approx. 1500).

4.2.8 Discussion and conclusion

4.2.8.1 Overall performance of DNN-HMM and transformer-based ASR systems on conversational speech

The main aim of this work was to gain insights about which aspects of casual, conversational speech cause the largest challenges for different ASR architectures. We conducted ASR experiments with four systems that are distinct with respect to

three aspects: **(A1)** HMM vs. transformer-based, **(A2)** amount of training data from the target language and style, and **(A3)** incorporation of explicit linguistic knowledge. In the following sections of this discussion, we point towards these aspects. The four different ASR systems were Whisper, Kaldi and wav2vec2 – with and without lexicon/LM (w2v and w2vLM). Whereas for Whisper, we performed no fine-tuning, as an example for a system not having any domain-specific data, the other architectures were informed with in-domain speech data.

In general, we found that all systems performed well on read speech (with speaker-dependent WERs of 11.8% (Whisper), 3.62% (Kaldi), 1.81% (w2v) and 1.01% (w2vLM)), in CS they all had troubles with specific conversations and did exceptionally well on others (overall WERs of 41.78% (Whisper), 42.86% (Kaldi), 29.81% (w2v) and 22.79% (w2vLM)). For the different conversations, we showed that Pearson correlation coefficients between conversation-dependent WERs were high ($> 60\%$ for Whisper vs. all and $> 87\%$ in case of Kaldi vs. wav2vec2). Our results show that the four ASR systems perform in the range of state-of-the-art for non-spontaneous speech material, they are however not robust in the task of recognizing casual conversational speech.

In the original paper of Whisper (Radford et al., 2023), the authors claimed that supervised speech recognition models trained entirely on (English) Librispeech (Panayotov et al., 2015) have very different robustness properties. This was demonstrated by the fact that Whisper out-performed earlier ASR results of benchmark Librispeech models on other English data sets (e.g., Common Voice (Ardila et al., 2020) or Switchboard (J. J. Godfrey et al., 1992)). Accordingly, they also claim that Whisper potentially complies with human behavior (they compared ASR errors with 95% confidence intervals of human errors), at least with respect to the results on English data sets. Our findings show a different picture with respect to robustness. Both, speaker-dependent and conversation-dependent means and standard deviations of WERs show a large range for Whisper (absolute difference between means and standard deviations were $\approx 30\%$ and 5.46%), and this despite the fact that the utilized Whisper system (large-v2) was also trained with large amounts of German speech data, for which Whisper even achieved better results than for English speech data (e.g., WERs of $5.5\% < 6.2\%$ (Multilingual Librispeech) or $6.4\% < 9.5\%$ (Common Voice 9)) **(Relates to A2)**.

Szymański et al. (2020) expressed their skepticism with respect to low WERs on benchmark data sets like, e.g., Switchboard (J. J. Godfrey et al., 1992) or Callhome. They found that on their internal multi-domain benchmarks, their ASR systems achieved WERs between 13.73% (for an insurance domain) and 22.16% (for booking and wireless telecommunication calls). Our findings further highlight this robustness problem of state-of-the-art ASR systems with respect to different unseen speakers or conversations. To conclude, we demonstrate that there are ongoing robustness problems of ASR systems that have not been fully resolved. This is one of the main reasons why we were driven to explore the causes of the challenges in recognizing conversational speech across different ASR architectures.

4.2.8.2 How WER is affected by utterance length and articulation rate

In our statistical analysis, we identified that the utterance length in number of **#tokens** and the articulation rate **AR** significantly affect utterance-level WERs in the ASR systems Whisper and Kaldi. The effect size of the quantitative interaction

of these two variables, as given by the Interaction-Forest EIM values, were the highest across all ASR architectures. When analyzing the overall distributions of the mean WERs at utterance level with respect to **#tokens**, we observed strong differences across ASR systems, especially when comparing Whisper with the other architectures which were trained or fine-tuned on in-domain data. For **AR** the picture was different. For utterances containing only one word token, the mean WERs on utterance level were usually best for lower articulation rates (with a slight exception in case of Whisper). In case of Whisper, a significant interaction term between **#tokens** and **AR** indicated that for *short* utterances (containing one to four word tokens) there is a *sweet spot* with good WERs for articulation rates between $10\,\mathrm{s}^{-1} - 12\,\mathrm{s}^{-1}$. For Kaldi, the WERs on utterance level tended to be worse as the number of tokens increased, but performance was less affected by articulation rate than for Whisper. For wav2vec2, we observed that for articulation rates up to $\mathbf{AR} \approx 12\,\mathrm{s}^{-1}$ WERs were better, but for $\mathbf{AR} > 12\,\mathrm{s}^{-1}$ the WERs were worse for w2v, but this effect was smaller for w2vLM. Thus the two systems having linguistic information in form of an LM, a pronunciation lexicon (Kaldi) or a simple word-level lexicon (w2vLM) resulted to be most robust against high articulation rates **(Relates to A3)**.

Hirschberg et al. (2004) found that HMM-based systems for human-machine interaction performed on average worse in longer turns (as measured in seconds) than in shorter turns. With respect to our utterance length feature **#tokens**, Kaldi achieved almost constant mean WERs for utterances of $2 \ldots 15$ tokens, indicating that next-generation hybrid DNN-HMM models are more robust than older GMM-HMM models with respect to utterance length. Not surprisingly, the same observation was also true for the up-to-date transformer-based wav2vec2 architecture **(Relates to A1)**. In contrast, (Wei et al., 2022) found that a transformer-based conversational ASR system achieved better recognition accuracy when including more contextual (historic) acoustic and linguistic information. Our results also show that more context, at least on utterance level, improved recognition results in case of the zero-shot Whisper system. With respect to tempo, Goldwater et al. (2008) analyzed how WERs on word level are affected by speech rate and reported little effects for words close to the average speech rate, but more errors for *"more extreme values"*. Interestingly, they also found fewest errors for words longer than the average. Their observation on speech rates is in line with our findings, where especially for Whisper, we found a *sweet spot* for the interaction of the features **#tokens** and **AR**.

To conclude, our analysis showed that utterance length and high articulation rate have a significant effect on ASR performance. Especially for *short* utterances, we revealed differences between the zero-shot architecture Whisper, which performed worse than Kaldi and wav2vec **(Relates to A2)**.

4.2.8.3 How WER is affected by the entropies of the RMS and F0 contours

In our study, we used the recently introduced pseudo-entropies (Linke, Kubin, & Schuppler, 2023) that describe the contour variation of F0 and RMS. We found that WERs on utterance level are highly correlated with the feature **HPSF0N** for *short* utterances (one to four word tokens), with the strongest effect observed for Whisper. Our statistical analysis revealed that WERs tended to be better for

single-word utterances of a rather uniform F0 contour (corresponding to high values of **HPSF0N**), and this effect was particularly strong for Whisper. For wav2v and Kaldi, the WERs were found to become more independent of the F0 contour, especially for utterances containing two to four word tokens, and this effect was strongest for w2vLM. We also found a significant qualitative interaction between **HPSF0N** and **AR**, which indicated best WERs for flat F0 contours in combination with slow speech. Note that the pseudo-entropy features might be less representative for very *long* utterances, as the acoustic relationships across words become more complex. Nevertheless, for half of the data (i.e., for *short* utterances), F0 and RMS (pseudo-)entropies explained important relationships across all ASR architectures, where especially Whisper's performance was affected by RMS (pseudo-)entropy.

Previous studies have not considered the entropy of F0 and RMS contours when analyzing ASR performance. For an HMM-based ASR system, Goldwater et al. (2008) found that *"more extreme values"* of pitch mean and pitch range were related to a higher WER. Those results are in line with our findings, as we showed that higher WERs tended to occur in utterances of less uniform F0 contours, or in other words, lower values of **HPSF0N**. Summing up, the performance of all ASR architectures was sensitive to F0 variation, especially with respect to single-token utterances. For utterances containing two to four word tokens, the performance of the systems trained or fine-tuned on domain-specific data was independent of the F0 variation **(Relates to A2)**.

4.2.8.4 How WER is affected by pronunciation variation

Given that the conversational speech material used was not only spontaneous but also spoken by speakers of a regional variety of German, our WER analysis focused also on gaining insights with respect to whether ASR performance is affected by how closely an utterance is spoken to the standard, canonical pronunciation of the words, as captured by the feature **PronLD**. In general, this feature correlated especially strongly with WER for Whisper, which is not surprising as this system did not see any in-domain data nor have access to a knowledge-based pronunciation lexicon (**Relates to A2**). Not surprisingly, for all systems, WERs on utterance level were best for *single-word* utterances that were produced closer to the standard pronunciation (low values of **PronLD**). For Whisper, we observed higher WERs for *short* and *long* utterances produced further away from the standard (high values of **PronLD**), whereas for the other systems WERs for *short* utterances were rather independent from **PronLD**. w2v (decoding without lexicon/LM) showed a different pattern, i.e., WERs were better for utterances with three and four word tokens and values of **PronLD** between approx. $1 - 2$. This result might be related to the transformer encoder of the wav2vec2 architecture which benefits from a large context **(Relates to A1)**. Furthermore, w2vLM (decoding with lexicon/LM) is slightly more robust against pronunciation variation as w2v **(Relates to A3)**. For Whisper, we further found that for utterances spoken far away from the standard pronunciation, especially when occurring at high values of **HPSF0N**, the WERs were worse compared to the other architectures that have been trained or fine-tuned on in-domain data **(Relates to A2)**.

Previous studies emphasized the impact of phonetic neighborhood density on HMM-based ASR. Goldwater et al. (2008) found that dense phonetic neighborhoods pose recognition challenges. However, our approach diverges as it uses CS from a

low-resourced variety. We aimed to gauge the deviation of utterance pronunciation from standard norms, an aspect overlooked in previous works. This approach is particularly relevant given that the majority of German speech models are trained on non-Austrian, often prepared speech. We hypothesized that utterances closer to standard pronunciation would be recognized more accurately. Our findings support this hypothesis, especially in the context of Whisper's low performance on utterances that were spoken with a pronunciation further away from standard German and find that a knowledge-based lexicon is beneficial on top of a transformer-based system.

4.2.8.5 How WER is affected by utterance-level perplexity

Compared to all other features analyzed, we found that **pplWIKI** had a generally weaker effect on the WERs on utterance level. With respect to the quantitative interaction between **#tokens** and **pplWIKI**, we found that WERs tended to be lower for *single-word* utterances and lower perplexities across all ASR architectures. Note that our analysis also showed that this quantitative interaction effect was even weaker with increasing values of **pplWIKI**. For utterances containing two to four word tokens, our analysis indicated that WERs generally were slightly worse with increasing perplexity. The (weak) qualitative interaction effect between **PronLD** and **pplWIKI** showed that across all architectures, utterances pronounced closer to the standard in 'not surprising' word sequences led to better WERs than for utterances spoken further away from the standard and having high values of **pplWIKI**. This result is as expected, especially from a human speech recognition point of view. For *long* utterances, we observed that WERs tended to be worse the longer they were, which also tended to have higher values of **pplWIKI**. Interestingly, this was also the case for wav2vec2, which however had higher absolute WERs on those utterances than the other systems.

Goldwater et al. (2008) found for an HMM-based system that there was an almost linear relationship between trigram-log-probabilities and WERs. In particular, higher values of the trigram-log-probabilities improved the results. This in line with our results, where lower perplexities (which is analogue to higher log-probabilites (cf. Eq. 4.15)) led to more or less better WERs. Although we also observed an effect of LM probabilities on the WERs, this effect was weaker than the effect of the other features (i.e., **#tokens**, **AR**, **HPSF0N** and **PronLD**), and this was the case for all ASR architectures. With respect to the calculation of the perplexity feature **pplWIKI**, we are aware of its potential limitations given the simple n-gram modelling approach. In subsequent work, we plan to incorporate neural LMs and investigate whether they better predict WERs.

4.2.8.6 How well do our results generalize to other corpora?

This study is based uniquely on conversational speech data from Austrian German, a non-dominant variety of the well-resourced language German. The question thus arises how well the findings reported here transfer to conversational speech data from other languages or language varieties. Can we expect similar findings for a conversational speech corpus of American English or of Scottish English? In order to allow, at least in principle, for a positive answer to that question, we designed our analysis such that style-specific and region-specific variation is captured in separate features. More precisely, we included four features for style-specific variation

(articulation rate **AR**, measures of variation for the F0 and RMS contours **HPSF0N** and **HPSRMSN**$_{Res}$, and language model perplexity **pplWIKI**) and one feature for region-specific variation (distance to the canonical pronunciation **PronLD**). The question whether our conclusions transfer to other conversational speech data thus receives evidence for an affirmative answer if i) the distribution of these features is similar to what we observed for GRASS and ii) the same features can be shown to have the same effects on ASR performances.

Addressing the first item, we note that the style of the corpus used here has the following characteristics: To name a few, utterance durations were $1.54\,\mathrm{s} \pm 1.42\,\mathrm{s}$, $\approx \frac{1}{3}$ of all utterances were single-word utterances, on average each word was spoken with $1.37 - 1.43$ pronunciation variants, and mean articulation rates were between $9 - 16$ phones per second. These numbers are not uncommon also for other corpora of conversational speech (for a comparison of GRASS with characteristica of other conversational speech corpora see Schuppler et al. 2017). At the same time, we observed that WERs vary across conversations (between 4% for w2vLM and 8% for Whisper), which indicates that each conversation creates its own conversational dynamic. Thus, while there will certainly be quantitative differences between the feature distributions of different corpora, we believe that the general, qualitative picture will be similar: namely, that lively, casual face-to-face conversations between two speakers are characterized by high articulation rate, short utterances, and substantial pronunciation variation. We thus claim that the respective feature distributions of GRASS are representative also of conversations in other languages or language varieties.

This leaves open the second item, namely whether the effects that certain features have on WERs will remain the same for conversations in other languages. Providing evidence for an affirmative answer to this question is more difficult. On a superficial level, such evidence is provided by the fact that our results are in line with the related literature. For example, with respect to language and style, many findings in (Goldwater et al., 2008) are still largely comparable to ours especially with respect to style-specific features like speech rates, measurements with respect to the F0 contour or LM perplexities, even though their analysis was based on American English telephone conversations. For a more detailed picture, one has to consider the same question for each ASR system separately. For example, at the moment, we have no reason to believe that our results transfer to other languages or varieties in the case of Whisper. Indeed, Whisper is trained on large amounts of speech data from multiple languages. The performance of Whisper – and hence also how this performance is affected by speech with a certain feature distribution – depends strongly on the amount and type of training data for the considered language. For example, the performance of Whisper could become independent of the pronunciation feature **PronLD** if its training data contained samples from the same variety as GRASS (or if it was fine-tuned on GRASS, cf. Sec. 4.2.8.8). Similarly, if a certain language subset of the training data covers a large range of articulation rates (e.g., because training data of this language contains a large portion of CS), the performance of Whisper for this language may not depend as strongly on **AR** as it does for Austrian German. At the other extreme, we believe that our results for Kaldi generalize very well across languages: The training process for Kaldi relies exclusively on data from the corpus under investigation. Hence, we have reason to expect that the trained Kaldi model has similar properties as

for GRASS, given that the respective corpus has a similar distribution of features. Wav2vec2 will, in our opinion, assume a position in the middle of the spectrum spanned by Whisper and Kaldi, respectively. Wav2vec2's self-supervised speech representations operate on a temporal scale that is shorter than words, which makes these representations *universal* for several languages and varieties. ASR systems based on wav2vec2 take these representations as input, and only the attention network (encoder) and projection layer (decoder) of the ASR system are fine-tuned using data from the corpus under consideration. And indeed, it was shown in Linke, Kadar, et al. (2023) that representations of CS of different varieties of the same language have similar distributions, indicating that our results for GRASS may carry over to CS of other German varieties. Since Linke, Kadar, et al. (2023) further showed that Hungarian CS leads to representations with a different distribution, we cannot conclude that our results for wav2vec2 carry over from Austrian German to other languages. This strongly suggests that future research is required that confirms – or rejects – the hypotheses about factors affecting ASR performance proposed in this work also for other languages.

4.2.8.7 Conclusion

In recent years, modern transformer-based architectures have shown impressive improvements for ASR, also for spontaneous speech. In this work, we presented ASR results for spontaneous conversations, which are characterized by quick turn changes (i.e., short utterances) and variety-specific, dialectal pronunciation. If we aim at making dialogue systems more social, where humans allow themselves to speak more naturally with the machine, excellent ASR performance on conversational speech is indispensable. Our analysis showed the importance to understand with which characteristics of conversational speech novel ASR architectures are struggling, and with which not. Earlier works have only analyzed how the performance of classical HMM-GMM-based systems is affected by lexical and prosodic characteristics of speech, thus this work fills the gap to gain insights with respect to **A1** comparing them to transformer-based architectures, **A2** gaining insights in the role of in-domain training data and **A3**, the role of linguistic knowledge incorporated into ASR systems. **Related to A2**, our analysis reveals that with zero-shot learning (Whisper), which has most probably not seen speech data similar to our corpus (i.e., spontaneous speech from a low-resourced variety of German), performance is poor, especially for *short* utterances and for large pronunciation variation, despite being trained on enormous amounts of data. Additionally, only the zero-shot system seems to be affected by the F0 and RMS contours, respectively, while all other systems seem to be quite robust against F0 and RMS variation. **Related to A1**, we found that transformer-based architectures that are fine-tuned on the speech corpus outperform HMM-based architectures that are trained on the same corpus by a large margin. For very short utterances, however, the HMM-based architectures perform well and (surprisingly) even outperform Whisper by far. For longer utterances, the perplexity of the word sequences plays a role too. This is particularly noticeable with wav2vec2, which tends to perform worse with higher perplexities and better with lower perplexities (yet, only few utterances were impacted by these values of mean perplexities). **Related to A3**, we observed throughout all experiments that the transformer-based architecture wav2vec2 combined with linguistic knowledge in form of a word-level language model and lexicon achieved the best performance and, what

is even more important, achieved highest robustness against acoustic and lexical variation. We thus see a great potential that for conversational speech produced in natural interaction, ASR will profit from a hybrid model constructed from a data-driven and a theory-driven component, including linguistic knowledge that goes beyond a simple lexicon (e.g., incorporation of pronunciation variants; knowledge about conversational dynamics and how the relate to syntactically differently formed utterance fragments).

4.2.8.8 Limitations

Our study has limitations that should be addressed in future research. First, we did not explore all possible combinations of system configurations as indicated in Tab. 4.1. For example, we cannot make a statement about how w2v fine-tuned on, e.g., a corpus of canonical German read speech performs in the CS setting of this study. Furthermore, even though the combination of a transformer-based ASR system without explicit linguistic knowledge that has been fine-tuned on data from the target language and style is covered by w2v, we admit that the peculiar performance of Whisper as shown in Fig. 4.5 suggests an analysis of Whisper fine-tuned on GRASS. Second, our experiments were limited to one corpus, which constrains the generalizability of our findings. However, we expect that our analysis raises the attention to those phenomena that are also relevant for other conversational speech corpora and that it motivates other researchers working on conversational speech from other languages and other dialects.

4.3 Conversational speech recognition needs data? Experiments with Austrian German

4.3.1 Motivation

Solving ASR tasks for conversational speech is crucial especially for social robots interacting with humans or automatic transcriptions of multimedia meetings (Popescu-Belis et al., 2012). Two humans who interact spontaneously with each other introduce complex inter- and intra-speaker variation depending on for instance the speaker's attitude towards the listener and the speaking task (Wright, 2006). Especially casual face-to-face conversations are characterized by a large amount of speaker-dependent pronunciation variation, by disfluencies, and by broken words or incomplete utterance structures. The resulting high degree of variation on all linguistic levels affects the acoustic model, the lexicon and language model of an ASR system.

Given the high variation in spontaneous conversations, the amount of annotated training data needed for ASR experiments to enable generalization for an unseen test set can sometimes be misleading in the sense that avoiding the data sparseness problem appears not to be possible, especially in case of spontaneous speech (Furui, 2009; Furui et al., 2005). Such studies give insights into the relationship between data size for acoustic model training and WER in case of Japanese spontaneous speech recognition: Utilizing 1/8 of available data (63.75 h) for acoustic model training results in a WER of approx. 27% whereby training with the entire data (510 h) gives an improvement of approx. 2%, but still no convergence.

This work deals with conversational speech from the GRASS corpus (cf. Sec. 2.2.1), which contains about 19 h (or 19 conversations) of Austrian German conversations introducing a considerable complexity in light of both inter-speaker and inter-conversation variation (i.e., from conversation to conversation, the amount of laughter, overlapping speech and disfluencies varies (Schuppler et al., 2017)). Despite German being a well resourced language, for the Austrian variety there are few resources available. For conversational speech, GRASS is the only resource currently available. For less spontaneous and less casual speaking styles, using German German[6] data for training an ASR system still delivers reasonably good results for recognizing Austrian German (Adda-Decker et al., 2013), this is, however, not the case for casual conversations where speakers show a higher degree of dialectal pronunciations. Hence, with respect to this variation, ASR experiments may require larger amounts of annotated conversational speech data than for less spontaneous speaking styles and thus may be viewed as a case of low-resourced (LR) language processing.

This section has been reformatted from:

[E] Julian Linke, Philip N. Garner, Gernot Kubin, and Barbara Schuppler. (2022). Conversational Speech Recognition Needs Data? Experiments with Austrian German. In *Proc. of LREC* (pp. 4684–4691).

My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing). Note that this section is based on the initial version of GRASS CS, in contrast to all other studies presented in this thesis, which were based on an updated version of GRASS with partial corrections of human annotations. However, this did not affect the data's comparability to other experiments, as the results described in this section align with all other findings.

[6]With German German, we refer to German as spoken by German speakers.

With wav2vec2.0 (Baevski, Zhou, et al., 2020), a framework for self-supervised learning of speech representations, powerful ASR models can be built also with small amounts of annotated data by fine-tuning pre-trained models. With the help of this modern architecture, it is even possible to come close to state-of-the-art results with only 10 min of labeled training data in the case of Librispeech (Baevski, Auli, & Mohamed, 2020; Conneau et al., 2021; Hsu et al., 2021; Panayotov et al., 2015; Zhang et al., 2021). Hence, we hypothesize this innovative framework also to be effective in solving a LR speech recognition task for Austrian conversational speech.

This study presents ASR experiments for Austrian German conversational speech from two ASR frameworks, the Kaldi speech recognition toolkit (Povey et al., 2011) and the wav2vec2.0 implementation of fairseq (Ott et al., 2019). In case of wav2vec2.0, we fine-tune a cross-lingual speech representation (XLSR) pre-trained model (Conneau et al., 2021) with different training data splits by testing each of the 19 GRASS conversations individually. Referring back to the problem of conversational speech complexity, we compare the XLSR experiments with an LR approach by pre-training and fine-tuning only with available GRASS conversational speech data. Ultimately, this study aims at investigating three hypotheses to gain more insight about the role of data for conversational speech:

1. Performing cross-validation by testing each conversation individually points out conversational speech complexity and reinforces a LR language processing assumption.

2. Fine-tuning a data-driven pre-trained cross-lingual speech representation model is effective for Austrian conversational speech.

3. Fine-tuning a LR speech representation model pre-trained only on Austrian conversational speech is not effective for Austrian conversational speech.

These hypotheses are investigated by the experiments presented in Sec. 4.3.3. After answering our hypotheses, the corollary in Sec. 4.3.4 discusses further findings which result from comparing the results from our ASR experiments.

4.3.2 Materials

GRASS corpus: The GRASS corpus (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017) contains about 19 h of Austrian conversational speech collected from 38 Austrian speakers (19f/19m). As language use in conversational speech varies strongly with educational level, social background and dialect region, GRASS contains only speakers who were born in the same broad dialect region (Eastern Austria), have been living in an urban area for years and have a higher education degree. For the conversational speech component, 19 pairs of speakers who had been knowing each other for several years were recorded for one hour each without interruption in order to encourage a fluent, spontaneous conversation. There was no restriction in terms of chosen topic or speaking behavior leading to the use of authentic, partly dialectal pronunciation with typical characteristics such as frequently occurring overlapping speech, laughter, or the use of swear words (Schuppler et al., 2017). No other person was present in the recording studio during the conversation. Despite the speakers' awareness of being recorded, they appeared to completely

Table 4.6: Summary of best and worst conversation-dependent WERs [%] (test set abbreviations include speaker IDs plus sex): Character-based (CHR) and phone-based (PHN) models with wav2vec2 are fine-tuned on LR pre-trained models (only GRASS) or XLSR. Kaldi models are also phone-based and incorporate additional pronunciations in the lexicon. We present results coming from 3 decoding strategies: Decoding without a lexicon (**Lexfree**), decoding with a lexicon (**Lex**) and decoding with a lexicon and LM (**4-gram**).

		Phone-based					
Kaldi	**Lexfree**	**Lex**	**4-gram**				
009M010M	-	-	65.12				
021F022F	-	-	43.89				
μ/σ	-	-	56.19/5.4				
					Character-based		
PHN-XLSR	**Lexfree**	**Lex**	**4-gram**	**CHR-XLSR**	**Lexfree**	**Lex**	**4-gram**
006M007M	-	42.03	32.71	006M007M	41.5	38.95	34.49
038F039F	-	26.63	17.44	038F039F	22.37	19.88	17.36
μ/σ	-	33.15/4.32	24.69/4.10	μ/σ	31.23/4.86	28.06/4.92	25.06/4.42
PHN-LR	**Lexfree**	**Lex**	**4-gram**	**CHR-LR**	**Lexfree**	**Lex**	**4-gram**
016M018M	-	90.44	73.45	016M018M	95.32	98.11	76.98
021F022F	-	64.93	45.14	038F039F	75.61	72.32	48.52
μ/σ	-	75.14/5.86	57.28/6.46	μ/σ	85.5/4.63	84.75/6.36	62.54/6.36

forget about the studio recording situation after a period of five to ten minutes, resulting in completely casual conversations.

Lexicon: All words from the GRASS corpus remaining after preprocessing are included in a lexicon file. For all phone-based experiments, we used the G2P online tool (Reichel & Kisler, 2014) for German German to create canonical German pronunciations, as a similar resource is not available for Austrian German.

Only for the Kaldi experiments, we derived additional pronunciation variants from the canonical pronunciations with 29 phonological rules based on findings from (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014). Some of the rules were assimilation and deletion rules relevant for conversational speech of all German varieties, whereas other rules cover pronunciations typical for the Austrian German variety. We added manually created pronunciation variants in order to capture specific pronunciations that cannot be generated in an automated way.

For wav2vec2 models, we create simplified lexicons where each word maps only to one pronunciation. In case of the character-based models, words are directly mapped to character sequences and in case of the phone-based models, words are directly mapped to canonical pronunciations.

4.3.3 Experiments

To investigate our three hypotheses, we first present experiments with Kaldi (cf. Sec. 4.3.3.1) and then experiments with fairseq (cf. Sec. 4.3.3.2).

4.3.3.1 Experiments with Kaldi

This section describes our experiments with Kaldi, which serve as a baseline for the main investigation.

Methods: When preprocessing GRASS transcriptions files for Kaldi, we excluded chunks involving artefacts, laughter and noise, resulting in a deletion of approx. 3.3 h of all available chunks ($\approx$ 17.5 h of all chunks from GRASS contain lexical items). In the end approx. 14 h of the data were used in the experiments.

We performed leave-p-out cross-validation (with $p = 2$ speakers of the same conversation) resulting in approx. 0.75 h of test data and 13.5 h of training data per split. Hence, we trained 19 baseline models (Kaldi) where each training split involves 18 conversations.

We reduced the initial phone set from 65 to 38 target phones by performing phone set minimization rules based on phonetic studies on Austrian German (Moosmüller, 2007): First, a replacement rule (silibant devoicing of the alveolar fricative /z/, as usual in Austrian German); second, a rule which split all diphthongs into two separate phones; third, a rule which united short vowels and long vowels.

We extracted 13-dimensional MFCCs and performed cepstral mean and variance normalization (CMVN). For acoustic models (AM), initial GMM-HMM-models comprised basic monophone and triphone training. On top of the triphone GMM, a speaker independent GMM model with linear discriminative analysis (LDA) and maximum likelihood linear transform (MLLT) (Gopinath, 1998) was trained. This model was the new basis for training with constrained maximum likelihood linear regression (fMLLR) (Gales, 1998). Finally, latter triphone alignments were used to train a TDNN with 13 layers and hidden dimensions of 512 by utilizing a cross-entropy criterion and only the existing MFCC features.

For Kaldi experiments, the language models (LM) were built using the SRILM toolkit with a Witten-Bell discounting for N-grams of different orders (Stolcke, 2002). LMs were trained on data coming from one training split. The experiments with 3-grams and 4-grams indicated a 4-gram model to be superior.

Results: With this Kaldi experiment, we aimed at investigating the hypothesis that testing each conversation individually points out conversational speech complexity and reinforces a LR language processing assumption. Tab. 4.6 shows the WERs achieved with our baseline Kaldi system. They range between 43.89% and 65.12%, where the resulting mean WER lies at approx. 56%, with a standard deviation of 5.4%. Hence, we observe a lack of performance and also high variation between the conversations with respect to the WERs.

The problem with conversational speech in LR scenarios is well-known: Results from Laurent et al. (2016) give WERs of $\approx$ 40% in case of conversational-like data. Sriranjani et al. (2015), for instance, showed that based on very limited LR Indian language data ($\leqslant$ 3 h) recorded in a rural environment WERs ranged from $\approx$ 10% to $\approx$ 34.5%. Furthermore, WERs from baseline experiments described in Yi et al. (2020) range from 33.77%...51.54% in case of LR multilingual telephone conversation data.

We find that performing cross-validation by testing each conversation individually points out conversational speech complexity and indicates a data sparsity problem. At this stage, we conclude that our first hypothesis cannot be rejected.

4.3.3.2 Experiments with Fairseq

This section describes our experiments with fairseq in order to further investigate the research questions.

Methods: In comparison to our Kaldi experiments, the preprocessing of GRASS transcriptions files was slightly different: We additionally had to exclude chunks involving foreign words and dialect lexemes, resulting in a total deletion of approx. 4 h of all available chunks (i.e., approx. 0.7 h more than for the Kaldi experiments). Other chunks which can involve breathings, speaker noise, singing, smacking, laughed speech, coughing, sighing, broken words or multi-word expressions were maintained. When parsing the transcriptions, we automatically corrected inconsistent orthography of fillers (e.g., `hm` and `hmm`), as these tokens can cause a high number of substitution errors. In the end approx. 13.5 h of the data remained for our experiments.

Just as in the Kaldi experiments, we perform cross-validation resulting in 19 training splits where each split results from leaving out 1 conversation. Subsequently, we receive approx. 0.75 h of test data and 12.75 h of training data per split. Finally, we randomly choose 10% of resulting training splits as validation sets (approx. 1.25 h) to adjust the LM weights in the decoder.

Training a speech recognition system with wav2vec2 involves two steps: 1) self-supervised learning from unlabeled speech data (pre-training) and 2) fine-tuning an obtained pre-trained model with labeled speech data. For all speech representation models, we used the same architecture with 315 Mparameters parameters containing 24 transformer blocks with model dimensions 1024, inner dimension 4096 and 16 attention heads.

When fine-tuning wav2vec2 models, we compared two basic target sets: 1) a character-based (CHR) set resulting in 31 characters as targets and 2) a phone-based (PHN) set resulting in 65 phonetic units as targets. Both target sets included a white space unit which models silence parts. Similar to our lexicon creation, in case of the character-based models, the orthographic transcriptions given by GRASS were directly mapped to character sequences. In case of the phone-based models, the orthographic transcriptions were mapped to canonical phonetic sequences.

The available pre-trained XLSR model was trained with 56 000 h of multilingual speech data built on top of wav2vec2. The training data of XLSR contains CommonVoice (36 languages, 3600 h) (Ardila et al., 2020), BABEL (17 languages, 1700 h) (Gales et al., 2014) and MLS (8 languages, 50 000 h) (Pratap et al., 2020). We fine-tuned XLSR with our labeled speech data with a CTC loss (Graves, Fernández, et al., 2006) after introducing a classification layer representing our targets. Here, we present results coming from 19 phone-based models (PHN-XLSR) and 19 character-based models (CHR-XLSR), as there are 19 conversations in GRASS.

For our experiments with LR wav2vec2 models, we pre-trained merely with in-domain GRASS data followed by fine-tuning the pre-trained GRASS models given the labels from our training splits. Also these models were trained with a Connectionist Temporal Classification (CTC) loss after introducing a classification layer representing the two target types. Thus, we view the resulting models as LR approaches, because we used exactly the same training data (11.5 h) for both, pre-training and fine-tuning. Also for this LR experiment with fairseq, we compare WERs coming from 19 phone-based (PHN-LR) and 19 character-based (CHR-LR) models.

For both XLSR and LR experiments with fairseq, we used a greedy decoder (**Lexfree**) in case of CHR models and a beam-search decoder without language model weighting (**Lex**) or with language model weighting (**4-gram**) in case of

CHR and PHN models. The greedy decoder searches the greedy best path by using only acoustic model predictions. The AM search space of the beam-search decoder is restricted by a lexicon and we incorporated an LM by providing an LM weight. Here, when incorporating an LM, we trained LMs of order 4 with modified Kneser-Ney smoothing and default pruning, which removes singletons of order 3 or higher by utilizing the KenLM toolkit (Heafield, 2011). LMs were trained merely on data coming from one training split and we choose an LM weight from the set of LM weights $\{1, 2, 3\}$ with respect to best WERs coming from the additional validation data. In case of beam-search decoding, we chose a beam size of 100. For the phone-based models, we do not provide results of the greedy decoder, because reasonable results could only be produced with the help of a lexicon introducing a target set which allows for word disambiguations.

Results from XLSR Pre-Training and GRASS Fine-Tuning: This experiment investigates the hypothesis that fine-tuning a data-driven pre-trained cross-lingual speech representation model is effective for Austrian conversational speech.

The middle row of Tab. 4.6 shows the WERs that achieved with the XLSR models. When decoding CHR-XLSR without a lexicon, WERs ranged between 22.37% and 41.5%, resulting in a mean value of 31.23% and standard deviation of approx. 5%. In case of PHN-XLSR, the **Lex** WERs are more similar to CHR-XLSR **Lexfree** results. **Lex** results of CHR-XLSR, on the other hand, are approx. 5% better with respect to mean value. We note that no big differences between **4-gram** PHN-XLSR and CHR-XLSR models can be observed, i.e., mean values and standard deviations are very similar. We observe that the powerful XLSR models give satisfactory results considering the high difficulty level of given face-to-face conversational data. As a matter of fact, all WERs of XLSR models are much lower than those from the Kaldi experiment, regardless of LM incorporation, and in case of CHR-XLSR even without utilizing a lexicon.

Wav2vec2 models pre-trained on 50 000 h of English data were tested on various languages showing their effectiveness in LR scenarios (Yi et al., 2020): Results with German telephone speech (approx. 13 h of training data) demonstrated an absolute improvement of approx. 20% compared to a baseline system. In general, they achieve more than 20% relative improvements in case of all six tested LR languages. Overall, their WERs are in the same range as ours.

We conclude that solving ASR tasks for GRASS conversational speech by fine-tuning speech representation models pre-trained on a high amount of out-of-domain data is effective. Thus, our second hypothesis cannot be rejected.

Results from GRASS Pre-Training and GRASS Fine-Tuning: This experiment investigates the last hypothesis that fine-tuning an LR speech representation model trained only on Austrian conversational speech is not effective for Austrian conversational speech.

The final row of Tab. 4.6 shows WERs achieved with the models PHN-LR and CHR-LR. The PHN-LR **4-gram** results were slightly worse than the results from the LR Kaldi approach, both with respect to mean WERs and to the worst conversation (i.e., by a difference of 8.33%). All PHN-LR models performed better than CHR-LR models, resulting in mean WER differences of 8.26% (**4-gram**) and

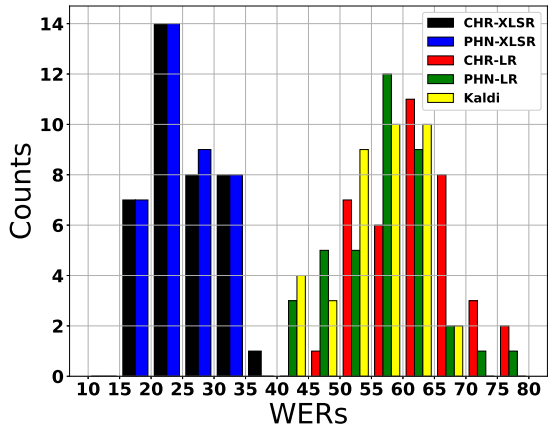


Figure 4.20: Histogram showing speaker-dependent WERs (**4-gram**). WERs and range of WERs are lower in case of fine-tuned XLSR models (35.71% . . . 16.09% = 19.62%) in comparison to fine-tuned LR models (79.37% . . . 43.36% = 36.01%). Kaldi model WERs range from 69.19% . . . 43.42% = 25.77%.

9.6% (**Lex**). Interestingly, **Lexfree** and **Lex** results were similarly bad in case of CHR-LR, with mean WERs of approx. 85%.

We refer back to Sec. 4.3.3.1 which presents WERs from the literature in case of LR conversational speech recognition, because the results from this experiment again demonstrate problems with respect to both, performance and robustness in case of LR scenarios.

From this experiment, we conclude that fine-tuning a LR speech representation model which is pre-trained merely on Austrian conversational speech is not effective. Also our Kaldi results demonstrate similar performance issues. Consequently, for neither of the two LR ASR approaches presented in this stduy, where models were trained merely on GRASS conversational speech, resulted in state-of-the-art WERs for conversational speech. Hence, our third hypothesis is true, and we show that training on approx. 11.5 h hours of conversational speech emphasizes the data sparsity problem. Additionally, these results are also reinforcing our first hypothesis, i.e., that performing cross-validation by testing each conversation individually points out conversational speech complexity and certifies the LR language processing assumption.

4.3.4 Corollary

After answering our hypotheses, this section discusses further findings which result from our experiments: we discuss the role of linguistic knowledge, the role of targets and the role of inter-speaker vs. inter-conversation variation.

Role of Linguistic Knowledge: We made several observations when looking at the influence of incorporating knowledge given by a lexicon or LM in case of wav2vec2 models.

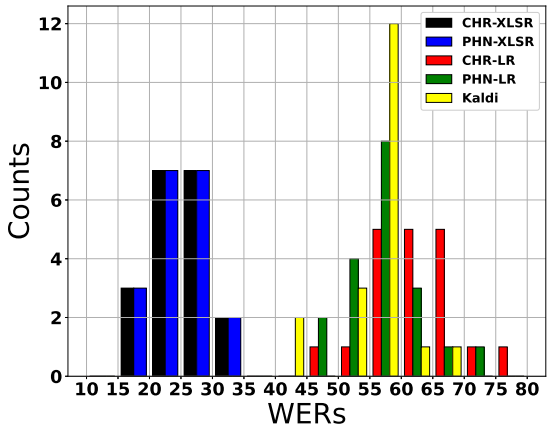


Figure 4.21: Histogram showing conversation-dependent WERs (**4-gram**). WERs and range of WERs are lower in case of fine-tuned XLSR models ($34.49\% \ldots 17.36\% = 17.13\%$) in comparison to fine-tuned LR models ($76.98\% \ldots 45.14\% = 31.84\%$). Kaldi model WERs range from $65.12\% \ldots 43.89\% = 21.23\%$.

Both, lexicon-based PHN/CHR-LR and PHN/CHR-XLSR models benefit from LM probabilites whereby higher differences in WERs can be observed in case of LR models ($\approx 20\%$ with respect to mean values). When comparing PHN-models with CHR-models those improvements are similar in the LR cases, but they differ more strongly in the XLSR cases, despite the overall WERs being similar. Hence, we notice that incorporating a LM has an higher impact on lexicon-based PHN models compared to lexicon-based CHR models in the XLSR case. At the same time, however, lexicon-free CHR-XLSR solutions are similar to lexicon-based PHN-XLSR solutions.

The experiments presented in Conneau et al. (2021) showed WER improvements of $\approx 2 \ldots 4\%$ due to LM incorporation when fine-tuning a smaller CHR-XLSR model. Another study showed improvements by adding LM probabilities and more advanced lexicons via dialect variation modeling (Khosravani et al., 2021). To the best of our knowledge, comparisons between PHN/CHR-XLSR models showing varying impacts of LM probabilities, have not yet been reported.

Looking at lexicon-based beam-search decoding results from PHN-XLSR and knowing that the AM search space is entirely restricted by the lexicon, one might argue that word mapping ambiguities lead to some substitution errors due to homophones. However, in case of our small canonical lexicon, only $\approx 1.8\%$ of all words are ambiguous introducing those unpredictable errors[7] and we believe that those errors are small in comparison to errors which arise from missing Austrian German pronunciations. Hence, we conclude that the canonical pronunciations in the lexicon, which introduce 65 target phones, lead to higher amounts of training errors in comparison to errors occurring from 31 character targets due to more noisy labels in case of Austrian German. We hypothesize this error to be lower in case of

[7]We hypothesize that, when introducing ambiguous pronunciations in the lexicon, words are randomly selected during beam-search decoding.

character-based systems, because they have only 31 character targets. Nevertheless, for both phone-based and character-based systems, incorporating LM probabilities resulted to help to reduce the impact of ambiguities.

Role of Targets: Here, we discuss the role of target labels by comparing our phone-based and character-based systems. If we look at performances from the LR models, character-based systems performed worse than phone-based systems in case of both, **Lex** and **4-gram**, whereby Kaldi WERs were more similar to PHN-LR **4-gram** WERs. XLSR models showed similar results when decoding with a LM, but in case of only lexicon-based decoding, CHR-XLSR models achieved better performances.

The systematic comparisons between character-based and phone-based ASR systems by Basson and Davel (2012) showed that increasing training data leads to similar performances in character-based and phone-based ASR systems. Zeineldeen et al. (2020) compared results for attention-based encoder-decoder models and found similar performances for character-based and phone-based systems regardless of lexicon or LM incorporation with more training data in general. Additionally, they also achieved similar results of 18.2% (PHN) and 18.6% (CHR) with a simplified decoder without LM nor lexicon by inserting word-disambiguate and end-of-word symbols in case of their phone-based models.

Our results are in line with results reported in the literature and suggest that character-based systems give similar performances as phone-based systems if enough data is available. However, our differences between phone-based and character-based models in case of lexicon-based decoding results indicate the relevance of knowledge, and that, for instance, the incorporation of more advanced lexicons might lead to further improvements.

Inter-Conversation vs. Inter-Speaker Variation: Our results indicate that variation in WERs with respect to each conversation and with respect to each speaker differs when comparing LR and XLSR models.

Tab. 4.6 shows that in case of beam-search decoding standard deviations of WERs are always higher in the LR scenario with wav2vec2 models than the WERs of the XLSR scenario. Fig. 4.20 clarifies this variation by comparing speaker-dependent WERs of **4-gram** models. In general, histograms over bins with 5%-width show that overall WERs and the range of speaker-dependent WERs are lower in case of XLSR models compared to LR models. Furthermore, when comparing WER ranges normalized by mean value, we measured values of 0.79 (PHN/CHR-XLSR), 0.6 (PHN/CHR-LR) and 0.46 (Kaldi). Fig. 4.21 clearly demonstrates that range of conversation-dependent WERs is lower in case of Kaldi models (21.23%) compared to wav2vec2 LR models (31.84%). In case of normalized conversation-dependent WER ranges, we measured values of 0.69 (PHN/CHR-XLSR), 0.53 (PHN/CHR-LR) and 0.38 (Kaldi). Corresponding entropy measurements which address directly to the shape of the distributions are 0.83 (PHN/CHR-XLSR), 0.96 (PHN/CHR-LR) and 0.79 (Kaldi). Even if absolute WER ranges of XLSR models are lowest, our measurements demonstrate that Kaldi distributions appear to have the least unexplained variability, especially in case of conversation-dependent WERs.

A broad study on domain shifts in self-supervised pre-training (Hsu et al., 2021) observe that adding more out-of-domain data during pre-training is beneficial and

simultaneously pre-training on more domains improves robustness in general.

Our findings confirm the effectiveness of fine-tuning GRASS conversational speech with XLSR with respect to performance, but we still observe lack of robustness with respect to resulting WER distributions. Additionally, in case of Kaldi models variation per conversation appears to be better modeled than variation per speaker.

4.3.5 Conclusions

In this work, we presented ASR experiments for Austrian German conversational speech from two ASR frameworks, the Kaldi speech recognition toolkit and fairseq (i.e., wav2vec2). We investigated the impact of data size, inter-speaker and inter-conversation variation, and structural knowledge for ASR performance, and compared phone-based and character-based ASR approaches.

Our results showed the effectiveness of fine-tuning a pre-trained cross-lingual speech representation model when solving LR ASR tasks with Austrian conversational speech. Even though performances were already satisfying with the data-driven approach, we still observed the importance of including structural linguistic knowledge via a lexicon or LM, as WERs decreased in case of both, LR and XLSR models. Furthermore, WERs varied strongly from speaker to speaker and from conversation to conversation, indicating the complexity of conversational speech, and also indicating the lack of robustness to speaker variation in case of all ASR approaches shown here.

In future, we will further investigate whether the impact of more advanced lexicons and LMs is larger for ASR of conversational speech in comparison to ASR of other less spontaneous and less casual speaking styles. Given our findings from this study, we hypothesize that better performing systems do not necessarily result in systems which are also more robust to inter-speaker variation.

4.4 What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers

4.4.1 Motivation

Research on ASR is strongly domain dependent due to the diversity of its applications (e.g., keyword spotting, dictation, and human interaction with social robots). Usually, each application is trained with different task-specific data sets. For continuous ASR, mostly two speaking styles are distinguished, read (RS) and spontaneous speech [8]. Probably the best-known RS corpus is Librispeech (Panayotov et al., 2015), where ASR performance already converges to its limit (2.5%) (Chung et al., 2021). Also for less spontaneous conversational speech (CS) (e.g., Switchboard corpus (J. Godfrey et al., 1992)), performance reaches benchmark limits (4.3%) (Tüske et al., 2021). Nevertheless, for more spontaneous CS (i.e., casual face-to-face conversations), performance ranges only between 16% and 33%, given high inter-speaker and inter-conversation variation Kim and Kang (2021); Linke et al. (2022).

One of the reasons for why ASR performance degrades with increasing degree of spontaneity is the reduced spectral space (Furui, 2009; Furui et al., 2005). The same authors also state that one of the most important research issues is how to train and adapt statistical models for speech recognition. Modern ASR architectures have a strong focus on adaptation by developing self-supervised learning of speech representations, such as those provided within the wav2vec2 framework, which make use of large amounts of unlabeled multilingual data (e.g., XLSR (Baevski, Zhou, et al., 2020; Conneau et al., 2021)). The experiments in Linke et al. (2022) and Khosravani et al. (2021) showed that ASR performance improves by fine-tuning the XLSR model with labeled data coming from a target domain (i.e., in both cases different varieties of German). For Hungarian conversational speech, Mihajlik et al. (2022) reached absolute WER improvements of approx. 12%. Furthermore, for telephone CS from low-resourced languages (BABEL) (Gales et al., 2014; Mary, n.d.), large WER improvements were reported on out-of-pre-training languages in comparison to baselines (e.g., absolute WER improvements of 9% on Swahili or 7.4% on Tagalog). The question arises what kind of information initial XLSR speech representations encode, as even out-of-pre-training languages seem to be well represented after fine-tuning. The aim of this work is to analyze initial XLSR speech representations to gain insights about how they encode data from different languages, their varieties, different speaking styles and different speakers. We aim at contributing to a better understanding of self-supervised speech representations, which is of interest not only to scientists in the field of ASR, but also to speech

This section has been reformatted from:

[F] Julian Linke, Mate Kadar, Gergely Dosinszky, Peter Mihajlik, Gernot Kubin, and Barbara Schuppler. (2023). What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers. In *Proc. of Interspeech* (pp. 5371–5375). My contribution roles were the conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization and writing (original draft and review/editing).

[8] Note that we further distinguish between more restricted conversational speech (e.g., telephone speech or task-oriented speech) and *casual face-to-face* conversations without any topical restrictions.

scientists interested in acoustic characteristics of different speaking styles.

Our approach towards finding an answer to this question is inspired by the analysis on similarity matrices of XLSR codebook entries for 12 or 17 different languages by Conneau et al. (2021). Whereas that study demonstrated how the codebook entries group together related languages, in this work, we take the approach one step further by analyzing not only different languages, but also different language varieties, and individual speakers of different speaking styles (i.e., read, spontaneous-task oriented, and casual conversational speech). More concretely, we compare two languages from to different languages families, where one is an out-of-pre-training language (i.e., Hungarian) and one is an in-pre-training language (i.e., German). In addition, we perform a speaker-wise analysis, allowing us not only to study the distances of languages, styles and varieties, but also the distances between speakers, as well as the distance of speakers with themselves when producing different styles. We aim at answering the more general research question of whether the frequency usage of shared discrete speech representations (given by XLSR) encode acoustic properties/characteristics for different languages, varieties, speaking styles and speakers.

4.4.2 Materials

The following experiments are based on German (G), Austrian German (AG) and Hungarian (H) corpora (cf. Tab. 4.7), covering read speech (RS) and conversational speech (CS) of *different degrees of spontaneity*: CS^+ for topic-free casual conversations and CS^- for task-oriented/task-restricted conversations. Sec. 2 provides a more detailed overview of the utilized speech corpora.

GRASS corpus: GRASS (Schuppler, Hagmüller, et al., 2014; Schuppler et al., 2017) contains 6 h of read (GRRS) and 19 h of conversational speech (GRCS) from 38 Austrian speakers (19f/19m). GRRS and GRCS are spoken by the same 38 speakers. For GRCS, 19 pairs of speakers who have known each other for several years were recorded for one hour and chosen topics were not restricted leading to casual speech. For the experiments with GRCS, chunks with artefacts, noise, whispering, foreign words and dialect lexemes were excluded, resulting in a total deletion of approx. 4 h, leaving approx. 13.5 h for our experiments. Then, filler labels were unified. We noticed long silence parts at the beginning of all GRRS chunks which could distort this analysis due to higher amounts of codebook usage relating to silence parts. Hence, we cut out 1.3 s of audio at the beginning of each file.

GECO corpus: The GECO corpus (Schweitzer et al., 2015) contains 46 spontaneous dialogues of approx. 25 minutes between female speakers. The corpus introduces settings GEMO with 22 dialogues, where participants were separated by a solid wall, and GEMU with 24 dialogues, with face-to-face conversations comparable to GRCS. In both settings, speakers were able to freely talk about any topic they want (thus classified as CS^+ in Tab. 4.7). For our experiments, GEMO and GEMU were preprocessed similar as GRCS and almost all chunks were kept.

KIEL corpus: The Kiel Corpus of Spoken German (KIEL) (Kohler et al., 2017) contains approx. 5 h of read and spontaneous speech produced by speakers from

Table 4.7: Overview of used data sets: Hungarian (H), German (G) and Austrian German (AG) corpora, containing read (RS) and conversational speech of different degrees of spontaneity (i.e., CS^+ for casual face-to-face conversations and CS^- for task-oriented/task-restricted conversations).

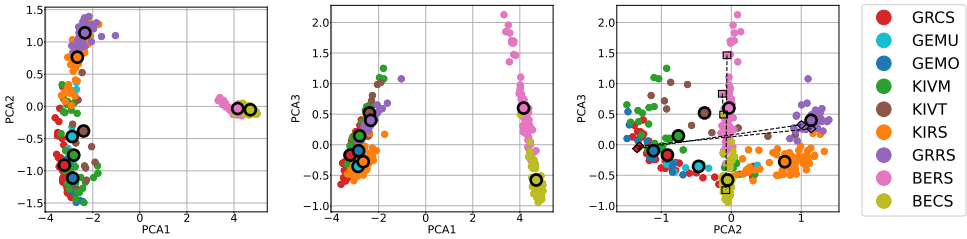
Corpus	**Abbr.**	**Style**	**Variety/ Lang.**	**Hours**
BEA Discourse	BECS	CS^-	H	14.2
BEA Readtext	BERS	RS	H	3.8
GECO-Multi	GEMU	CS^+	G	9.8
GECO-Mono	GEMO	CS^-	G	8.92
GRASS CS	GRCS	CS^+	AG	13.5
GRASS RS	GRRS	RS	AG	4.6
KIEL-Verbmobil	KIVM	CS^-	G	3.72
KIEL-Videotask	KIVT	CS^-	G	1.3
KIEL RS	KIRS	RS	G	2.8

Northern Germany. The read speech (KIRS) contains sentences and stories from 53 speakers (26f/27m). KIVM contains approx. 4 h of dialogues from 43 speakers (22f/21m) who were making appointments and KIVT contains approx. 1 h of dyadic conversations. In As KIVM and KIVT contain task-oriented/topic-restricted dialogues, we classify them as CS^- in Tab. 4.7. As for GRCS, also for KIVM and KIVT chunks with laughed speech and noise were excluded and filler annotations were unified. For KIRS, depending on given transcription material, we utilized already trimmed audio-files directly or trimmed the audio-files on the basis of the boundary markers of given annotations. In case of all GECO and KIEL corpus components, we excluded resulting chunks with durations greater than 20 s due to our limited computational infrastructure.

BEA database: The original BEA database ("BEszélt nyelvi Adatbázis" in Hungarian, meaning spoken language database) aimed at collecting studio quality speech data from 500 speakers (Neuberger et al., 2014). For the experiments, we used the BEA-Base subset (Mihajlik et al., 2022) of the database, specifically the *Readtext* (BERS) and *Discourse* (BECS) modules of the "train-114" subset. Both, BERS and BECS included the same speakers while female and male participants were closely balanced. In case of BECS, each conversation was recorded approx. 45 min and one experimenter guided the casual conversations between the speaker and an optional discourse partner on various random topics. Conversations from BECS included recordings which relate to only one speaker which makes it possible to compare specific speakers between BECS and BERS but, different from GRCS, it is impossible to compare one speaker pair from BECS with respective speakers from BERS.

4.4.3 Analysis of self-supervised speech representations

We hypothesize that shared discrete speech representations of different corpora encode speaking styles and varieties. Here, we investigate this hypothesis by analyzing similarity matrices resulting from a comparison of normalized frequency usage of



(a) PCA1 and PCA2 (b) PCA1 and PCA3 (c) PCA2 and PCA3 (d) Legend

Figure 4.22: Speaker-dependent codebook usage with respect to the considered German and Hungarian corpora in the 3-dimensional PCA space after transforming their similarity matrix which results from codebook frequency usage of XLSR. BECS (olive) and BERS (pink) as well as GRCS (red) and GRRS (purple) involve the same speakers and filled circles with black outlines indicate corpus centroids. Dashed line connections of black rectangles and diamonds in (c) illustrate distances between BECS and BERS referring to same speakers as well as GRCS and GRRS referring to same speakers of one GRCS conversation.

discrete XLSR speech representations (introduced by codebooks) from different data sets. The source code related to our analysis is publicly available and can be accessed on GitLab[9].

From similarity matrix to PCA space: We used wav2vec2 (Baevski, Zhou, et al., 2020) with fairseq (Ott et al., 2019) to compute discrete shared speech representations with a multilingual pre-trained model (XLSR) (Hsu et al., 2021). XLSR is pre-trained in self-supervision with 56 000 h of speech data coming from 53 languages including German but not Hungarian and comprising approx. 99% of read speech and 1% of spontaneous speech (BABEL). XLSR has 315 M parameters containing 24 transformer blocks with model dimensions 1024, inner dimension 4096 and 16 attention heads. Given the pre-trained model, we computed latent speech representations and utilized the model's quantizer to obtain respective codebook indices of shared discrete representations. The quantizer is based on product quantization introducing $G = 2$ codebooks with each of them having $V = 320$ entries, resulting in a total number of 102400 possible codebook combinations.

In order to compare the frequency usage of speech representations coming from XLSR with respect to speakers, we quantized the utterances of each speaker of each preprocessed corpus and counted the utilized codebook entries. Then, we normalized each speaker's frequency usage with the total number of features per speaker, resulting in speaker-dependent prior distributions of codebook usage. Given these priors, we generated a similarity matrix by computing similarities of resulting distributions with a Jensen-Shannon divergence. Finally, the similarity matrix was transformed to a 3-dimensional PCA space.

Interpretation of three PCA dimensions: Fig. 4.22 shows 3 speaker-dependent scatter plots (PCA1/PCA2, PCA1/PCA3 and PCA2/PCA3) from the resulting 3-dimensional PCA-space. Speakers of each corpus are depicted in a different color. First thing, we notice is that PCA1 describes language, where component

[9]https://gitlab.tugraz.at/speech/speechcodebookanalysis

values > 0 categorize Hungarian speech and component values < 0 (Austrian) German speech. Second thing, we notice is that PCA2 separates the same GRASS speakers in different clusters based on speaking style. In general, we observe that PCA2 characterizes our degree of spontaneity within (Austrian) German where components > 0 visualize almost non-overlapping RS corpora. In the opposite direction, conversational components of higher spontaneity may overlap. Third thing, we notice is that PCA3 distinguishes Hungarian speaking styles where components > 0 define Hungarian read speech and components < 0 Hungarian conversational speech.

Centroids and their distances: At first, we compare resulting Hungarian centroids and (Austrian) German centroids in the 3-dimensional PCA-space (cf. filled circles with black outlines) with respect to Euclidean distances. In case of Hungarian centroids, we measured a distance of 1.3 between BECS and BERS which is mainly described by PCA3. In order to gain more insights into how speech representations differ between BECS and BERS, we randomly selected two speakers within BECS and measured their Euclidean distance to BERS resulting in 2.47 and 0.35 (cf. black dashed lines between olive and pink diamonds in Fig. 4.22). In general, mean and standard deviation of distances between same Hungarian speakers were 1.3 ± 0.7. In case of (Austrian) German centroids, we compared the resulting centroid of GRCS with the other 6 German-speaking centroids. We observe the smallest Euclidean distance between GRCS and GEMO (0.46) followed by distances with KIVM (0.53), GEMU (0.58), KIVT (1.19) and KIRS (1.77). The highest distance was between GRCS and GRRS (2.3), which is to some extent surprising as these two corpora contain speech from the same speakers. In order to gain more insights into how speech representations differ between GRCS and GRRS, we measured the Euclidean distance between a speaker pair within GRCS and to GRRS. We find that their distance in GRCS is approx. 0.07, whereas distances between the same speaker in GRCS and GRRS are considerably higher, i.e., approx. 2.56 and 2.4 (cf. black dashed lines between red and purple rectangles in Fig. 4.22). In general, mean and standard deviation of distances between same Austrian German speakers were 2.3 ± 0.4. Overall, when comparing the distances of all 19 speaker pairs, we found no correlation between GRCS and GRRS (Pearson Correlation Coefficient: $r \approx 0.02$, $p \approx 0.93$). These results show that the speech representations are more sensitive to the speech characteristics typical for read vs. conversational speech than to speaker specific characteristics. Finally, we compared the resulting centroid of BECS with German-speaking centroids and resulting centroid of GRCS with Hungarian-speaking centroids. We observe high Euclidean distances > 7.1 between BECS and the 6 (Austrian) German centroids with the smallest distance to KIVT (7.18) and the highest distance to GRCS (7.96). Overall, distances between GRCS and Hungarian centroids were > 7.4 since distance to BERS was 7.46.

Clustering of the 3-dimensional PCA space: Next, we performed k-Means clustering by using the resulting 3-dimensional PCA space with 6 clusters. This clustering enables classification by evaluating Euclidean distances to the 6 generated cluster centroids. We only measured the 2-dimensional distances with respect to projections in PCA2/PCA3, because those dimensions describe (Austrian) German (PCA2) and Hungarian (PCA3) speaking styles which is the focus of this study.

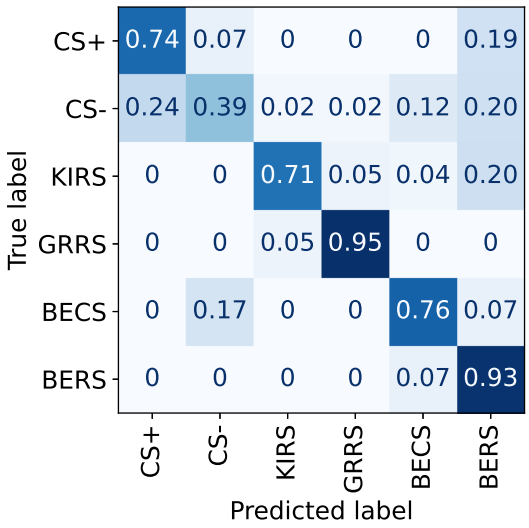


Figure 4.23: Resulting confusion matrix when clustering the 3-dimensional PCA space of the speaker-dependent similarity matrix (cf. Fig. 4.22) with k-Means introducing 6 centroids.

Fig. 4.23 shows the resulting confusion matrix. The clusters correlate with the degree of (Austrian) German spontaneity (CS^+ and CS^-), correlate for (Austrian) German read speech with variety (GRRS and KIRS) and for Hungarian speech with speaking style (BERS and BECS). Interestingly, for German, clustering did not separate variety, but only the degree of spontaneity.

With respect to the confusions that occur, nearly all speakers from both (Austrian) German RS corpora were assigned correctly (KIRS: approx. 70%; GRRS: 90%), whereas only approx. 40% of speakers from CS^- corpora were correctly assigned as CS^-, while approx. 20% of them were confused with CS^+, 2% of them were confused with KIRS and GRRS, 10% of them were confused with BECS and 20% of them were confused with BERS. In general, confusions of CS^+, CS^- and KIRS with BEA (approx. 20% in case of BERS) can be explained by our analysis approach which compares only distances within the dimensions PCA2 and PCA3[10]. Likewise, assigning speakers from speaking style CS^+ was easier in general leading to a confusion with CS^- of only approx. 7%. F1-scores of CS^+ and CS^- were 0.77 and 0.45. In case of BECS approx. 80% of the speakers were correctly assigned, while approx. 20% of them were confused with CS^- and approx. 7% of them were confused with BERS. Likewise, in case of BERS approx. 80% of the speakers were correctly assigned, while approx. only 7% of them were confused with BECS. Corresponding F1-scores of BECS and BERS were 0.78 and 0.74. These clustering results are in line with our earlier observation, as there is no confusion between GRCS (CS^+) and GRRS. Simultaneously, confusions between Hungarian speaking styles, namely BECS and BERS, were also small.

[10]Note that we could easily implement a condition on PCA1 if the aim of our study would be a better performing classification task

4.4.4 Discussion and conclusion

The main aim of this work was to test the hypothesis that shared discrete speech representations from speakers of different corpora encode languages, varieties and speaking styles. To analyze this hypothesis, we performed a clustering experiment with XLSR codebook entries from the different data sets, demonstrating that, in addition to languages, read and spontaneous speaking styles are indeed also distinguished in this feature space. Based on a 3-dimensional PCA space, independent of language (PCA1) almost all speakers from the read speech corpora were assigned correctly to the corresponding clusters, for the spontaneous corpora, however, this was only the case with CS^{+} and BECS with corresponding F1-scores of 0.77 and 0.78. We observed that speech representations of German spontaneous speaking style showed variety-independence, which we explain by the strongly varying speech representation usage. For read speech, we can distinguish between the German and Austrian German variety. In general, our findings are in line with those in the literature: The study by (Conneau et al., 2021) used similar methods to cluster discrete speech representations of multilingual pre-trained wav2vec2 models, demonstrating the possibility of grouping related languages. Another study on dialect clustering with sentence vector representations based on character-based metrics also generated plausible clusters (Sato & Heffernan, 2020). They found three emerging noticeable clusters in case of Japanese varieties, namely Tohoku dialect, Tokyo dialect and a combination of three Western dialects (Kansai, Chugoku and Kyushu).

Another focus of our analysis was on how the speech representations of the same speakers behave and whether they explain different degrees of spontaneity. We found that Austrian German speakers differ the most between different styles since mean distance of same Austrian German speakers was high (2.3). In contrast, mean distance of same Hungarian speakers was smaller (1.3). Furthermore, we found that Austrian German speakers also differ more from themselves within different styles, indicating speaker identity independence of the speech representations. Overall, our results indicate that speech representations vary the most among Austrian German speakers. Also (Asami et al., 2014) found that GMM supervectors based on utterances can discriminate read and spontaneous speech with less speaker-dependency. Simultaneously, the authors state that clustering spontaneous utterances is more difficult than read utterances.

To conclude, the results suggest that distance calculation based on shared quantized latent speech representations is also meaningful on a much finer granularity level (i.e., per speaker per speaking style) than it was introduced in Conneau et al. (2021) for languages. This may open new perspectives in speech data selection both for supervised and self-supervised learning, as speech sections matching the desired development set (or speaking style) could be collected at a relative low cost, requiring only a pre-trained wav2vec2 model but without the need of any additional information beyond the waveform. Furthermore, it may be worth exploring meaningful acoustic correlates that could shed more light on the nature of elusive self-supervised speech representations. We are going to extend our investigations in these directions in the future.

4.5 Prominence-aware automatic speech recognition

4.5.1 Motivation

Prominence classification experiments (cf. Chap. 4.5) have demonstrated the ability to distinguish between prominent and non-prominent words in conversational speech, highlighting word duration as the most important feature. However, these approaches currently require word alignments to allow the training of classification models. Additionally, the ASR results and analysis as described in Sec. 4.2 indicate that prosodic features impact ASR performance with respect to different ASR architectures. This section explores a novel approach that combines prominence detection and speech recognition by training a prominence-aware ASR system. First, we investigate whether automatic prominence detection by means of a fine-tuned transformer-based system achieves a performance in the range of the inter-annotator agreements. Second, we explore the potential of integrating prosodic information in terms of word-level prominence levels into ASR systems. The incorporation of prosodic information into ASR systems opens up new possibilities for the development of future applications, particularly for linguistic annotations and for prosody-informed dialogue systems.

4.5.2 Prominence Detection

4.5.2.1 Materials and Methods

Data preparation: For training and testing of the prominence detectors, we used data from the prosodically annotated subset of GRASS CS. The prominence-annotated subset includes prosodic annotations created by phonetically trained transcribers for a total of 4944 utterances including 15664 word tokens from 34 speakers (cf. Sec. 3.3)[11]. The prominence annotations distinguished the prominence levels 0 (no prominence; *PL0*), 1 (weak prominence; *PL1*), 2 (strong prominence) and 3 (emphatic prominence). Prominence levels 2 and 3 were combined as *PL2*.

Prominence detection: Prominence detectors were developed by fine-tuning the wav2vec2 XLSR model (Baevski, Zhou, et al., 2020; Conneau et al., 2021) (cf. Sec. 4.2 and 4.3) with the prominence-annotated utterances and a CTC loss (Graves, Fernández, et al., 2006). More precisely, we trained two separate prominence detectors PDET_{02} and PDET_{012}, where the first detector classified two prominence levels (*PL0* vs. *PL2*) and the second detector classified three prominence levels (*PL0* vs. *PL1* vs. *PL2*). Tab. 4.8 gives an overview of the data with respect to the two types of models. The reference text for training included only the resulting prominence levels as single numbers plus word boundary markers ("|"). Note that prominence annotations referred to prosodic words (e.g., the prosodic word "| sie hat |" was annotated as *PL0*). For a more detailed view of available transcriptions,

[11] Note that there was more data available for this experiment in comparison to the prominence classification experiment as described in Sec. 3.3. This is because the current experiment extracts features directly from raw audio data while the prominence classification experiment relied on a feature extractor for more specific prosodic features (F0, RMS and DUR) which occasionally resulted in missing values.

Table 4.8: Overview of the used Austrian German speech data for the prominence detectors $PDET_{02}$ and $PDET_{012}$. The table shows orthography and corresponding reference examples while the prominence detectors were exclusively trained using the references. The table also indicates the number of utterances (**#utts**) and the mean number of tokens plus standard deviation (**#tkns**).

Type	**Orthography**	**Reference**	**#utts**	**#tkns**
$PDET_{02}$	| sie hat | erzählt |	| 0 | 2 |	1770	2.09 ± 1.39
$PDET_{012}$	| wah | voll | nett |	| 0 | 2 | 1 |	4944	3.17 ± 2.13

Fig. 4.24 shows corresponding forced alignments and human annotations of the utterance with orthography "| sie hat | erzählt |". For each type of detector, we performed 10-fold cross-validation in order to test the generalization ability of the prominence detectors and provide corresponding accuracy means and standard deviations. Additionally, we trained models for one held-out test conversation (i.e., conversation with ID 004M024F). For evaluation, we compare 1) prominence detection error rates (PER) calculated similarly as word error rates while considering only prominence levels and word boundary markers and 2) accuracies, F1-scores and recalls for prominent words but only if an alignment between human-annotated word boundaries and detection-annotated word boundaries was possible with respect to each utterance.

In a final step, the entire GRCS component was automatically annotated twice with the final prominence detectors $PDET_{02}$ and $PDET_{012}$. For each utterance, if the detection results aligned with the word boundaries of given forced alignments of a Kaldi system (cf. Sec. 2.3.2)[12] only these words were automatically annotated with a prominence level (i.e., with respect to each speaker approx. $52.06\% \pm 8.57\%$ ($PDET_{012}$) and $42.3\% \pm 8.4\%$ ($PDET_{02}$) of the utterances were aligned). These automatically annotated words were then utilized as additional information for prominence-aware ASR training.

4.5.2.2 Results for prominence detection

Tab. 4.10 shows prominence detection results for all types of models. For $PDET_{02}$ we achieved PERs of $24.83\% \pm 1.79\%$ (10-fold CV) and 29.58% (004M024F). For this model, it was possible to align $69.56\% \pm 3.00\%$ (10-fold CV) or 63.48% (004M024F) of the utterances with respect to the detected word boundaries. For these words, we achieved accuracies of $89.72\% \pm 3.26\%$ (10-fold CV) or 87.40% (004M024F).

In contrast, for $PDET_{012}$ we achieved worse PERs of $36.54\% \pm 0.92\%$ (10-fold CV) and 41.02% (004M024F). This time, it was possible to align $66.80\% \pm 1.66\%$ (10-fold CV) or 64.34% (004M024F) of the utterances with respect to the detected word boundaries. Furthermore, we achieved worse accuracies of $69.45\% \pm 2.11\%$ (10-fold CV) or 64.97% (004M024F).

Confusion matrices in Fig. 4.25 illustrate in more detail results for conversation with ID 004M024F. With respect to recalls of $PDET_{02}$ (for 119 aligned words out of 73 utterances), it can be seen that 84% of *PL0* were correctly classified as *PL0* and 87% of *PL2* were correctly classified as *PL2*. Respective F1-scores were 83%/88% (*PL0*/*PL2*). For $PDET_{012}$, recalls (for 451 aligned words out of 184 utterances)

[12]For consistency, the automatic annotation of the entire GRASS CS component was based on word boundaries coming from forced alignments as human-annotated word boundaries are only available for the smaller prominence-annotated subset.

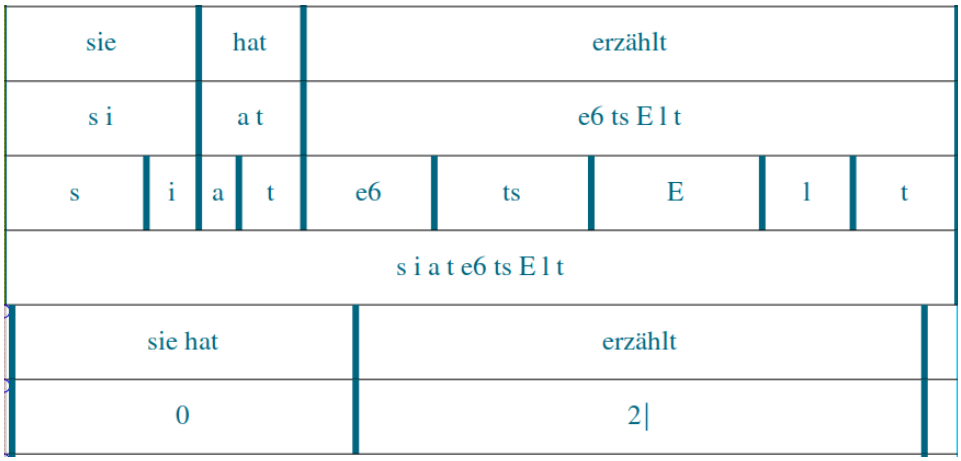


Figure 4.24: Transcriptions of one utterance from the prominence-annotated subset with forced alignments (rows 1-4) and human annotations (rows 5-6). The first four rows illustrate the output of the forced alignments, including: 1) word alignments with word text, 2) word alignments with realized pronunciations, 3) phone alignments, and 4) realized pronunciation without any boundaries. The subsequent two rows represent human annotations: 5) prosodic words with boundaries and 6) corresponding stress levels with boundaries. This example shows that boundaries of the prosodic words can be different from word boundaries of the forced alignments: The first prosodic word "| sie hat |" was annotated as *PL0*, while the second prosodic word "| erzählt |" was annotated as *PL2*. Note that the stress label "2|" indicates that the word was also annotated with a flat pitch contour.

of *PL0*/*PL2* were worse with 79%/62%. There were also strong confusions with respect to *PL1* where only 49% of *PL1* were correctly classified as *PL1* but 30% as *PL0* and 21% as *PL2*.

For conversation with ID 004M024F, it was also possible to evaluate prominence detection results with respect to the human-annotated labels by keeping only the prominence level information plus word boundary markers in the hypothesis text of the **Lexfree** models[13]. This results in worse PERs of 65.42%/73.52% for ASR02(PDET02)/ASR012(PDET012) compared to PDET02/PDET012, partly because not every hypothesis necessarily contains prominence information. This is also reflected in the quality of the alignments for which only 52.17%/43.01% ($\text{ASR}_{02}(\text{PDET}_{02})/\text{ASR}_{012}(\text{PDET}_{012})$) of the utterances were aligned with respect to word boundaries. Nevertheless, the accuracies of 85.53%/64.57% ($\text{ASR}_{02}(\text{PDET}_{02})/\text{ASR}_{012}(\text{PDET}_{012})$) showed comparable results to the original prominence detection models.

4.5.3 Prominence-aware ASR

4.5.3.1 Materials and Methods

Data preparation: Prominence-aware ASR systems were based on labeled speech data from the entire GRASS CS component. Pre-processing involved the exclusion

[13]More precisely, prominence levels were assigned bymajority voting of strings between word boundaries (e.g., the hypothesis "|d0 i0 e0|" becomes the string "000" which was assigned as *PL0* but the hypothesis "|d0 i1 e|" becomes the string "01" which was assigned as an empty string because no clear assignment of a prominence level can be made due to the ambiguity).

Table 4.9: Concept of character-based prominence-aware ASR training. Generally, each character in the reference text was assigned with a detected prominence level if possible or desired. ASR systems based on PDET_{02} allow training with a maximum number of two prominence levels (i.e., leading to the systems $\text{ASR}_0(\text{PDET}_{02})$, $\text{ASR}_2(\text{PDET}_{02})$ and $\text{ASR}_{02}(\text{PDET}_{02})$). ASR systems based on PDET_{012} allow training with a maximum number of three prominence levels (i.e., leading to the systems $\text{ASR}_0(\text{PDET}_{012})$, $\text{ASR}_2(\text{PDET}_{012})$, $\text{ASR}_{02}(\text{PDET}_{012})$ and $\text{ASR}_{012}(\text{PDET}_{012})$).

Type	Orthography	Reference
$\text{ASR}_{02}(\text{PDET}_{02})$	| die | waren | alle |	| d0 i0 e0 | w a r e n | a2 l2 l2 e2 |
$\text{ASR}_{012}(\text{PDET}_{012})$	| die | waren | alle |	| d0 i0 e0 | w1 a1 r1 e1 n1 | a2 l2 l2 e2 |

of utterances containing laughter, singing, imitations/onomatopoeia, unintelligible word tokens and artefacts which resulted in approx. 14.4 h (relating to 33734 utterances) of CS data. We standardized typical backchannels to `mhm`, removed punctuation marks and standardized the text to lowercase (cf. Sec. 4.2.3).

Prominence-aware ASR: For ASR, we again fine-tuned the pre-trained XLSR model (Baevski, Zhou, et al., 2020; Conneau et al., 2021) with a CTC loss (Graves, Fernández, et al., 2006) but in this case with additional information of prominence levels derived from the prominence detectors PDET_{02} and PDET_{012}. Tab. 4.9 shows how the automatic annotations were incorporated into the character-based models by modifying the reference text. For ASR systems based on automatic annotations from PDET_{02}, we trained models which include

- only prominence level *PL0* (ASR_0 with ≈ 69 character tokens[14]),
- only prominence level *PL2* (ASR_2 with ≈ 69 character tokens[14]),
- or both prominence levels *PL0/PL2* (ASR_{02} with ≈ 102 character tokens[14]).

For ASR systems based on automatic annotations from PDET_{012}, we trained models which include

- only prominence level *PL0* (ASR_0 with ≈ 69 character tokens[14]),
- only prominence level *PL2* (ASR_2 with ≈ 69 character tokens[14]),
- two prominence levels *PL0/PL2* (ASR_{02} with ≈ 102 character tokens[14]),
- or all three prominence levels *PL0/PL1/PL2* (ASR_{012} with ≈ 134 character tokens[14]).

For decoding, we used a greedy decoder (**Lexfree**) and a beam-search decoder with (**Lex**) and without language model weighting (**3-gram**). We used the same lexicon for all models by simply mapping all GRCS words to their character sequences. The 3-gram LMs were trained with data from each training split with the KenLM toolkit (Heafield, 2011) by using modified Kneser-Ney smoothing and default pruning. We evaluated ASR results on two conversations, namely conversation with ID 003M023F (which was not part of the prominence-annotated subset) and conversation with ID 004M024F (which was also part of the prominence-annotated subset). All ASR results are compared to a wav2vec2 baseline as described in Sec. 4.2.4 (≈ 37 character tokens[14]).

[14]Note that the number of character tokens can vary with respect to a given training set.

Table 4.10: Prominence detection results of prominence detectors $PDET_{02}$ and $PDET_{012}$ for two test conditions. The prominence error rates (**PER**) [%] and accuracies [%] of 10-fold CV results are shown with mean and standard deviations. The **PER** was calculated for all utterances of a test split. The ratio of possible alignments given correct word boundaries of an utterance for each test split (**%Aligned**) explains for which amount of utterances the word-level accuracy measurements could be calculated (**Accuracy**).

Type	Test set	PER	%Aligned	Accuracy
$PDET_{02}$	10-fold CV	24.83 ± 1.79	69.56 ± 3.00	89.72 ± 3.26
	004M024F	29.58	63.48	87.40
$ASR_{02}(PDET_{02})$	004M024F	65.42	52.17	85.53
$PDET_{012}$	10-fold CV	36.54 ± 0.92	66.80 ± 1.66	69.45 ± 2.11
	004M024F	41.02	64.34	64.97
$ASR_{012}(PDET_{012})$	004M024F	73.52	43.01	64.57

Figure 4.25: Confusion matrices derived from prominence detectors $PDET_{02}$ and $PDET_{012}$ for conversation with ID 004M024F. Results refer only to words of utterances where alignment between human-annotated word boundaries and detection-annotated word boundaries was possible.

4.5.3.2 Results for prominence-aware ASR

Tab. 4.11 shows resulting WERs of a baseline and prominence-aware ASR systems for conversations with IDs 003M023F and 004M024F. For the baseline experiments without prominence information (cf. Sec. 4.2.4), WERs ranged between $18.57\% - 26.04\%$ (003M023F) and $23.71\% - 31.25\%$ (003M023F)[15]. In general, WERs of prominence-aware ASR systems were worse than the baseline systems with absolute maximum deterioration of $2.1\% - 2.3\%$ in case of $ASR_{02}(PDET_{02})$ and **Lex**. An exception was the WER of ASR_0(PDET012), which was better than the baseline at 18.23%, but this improvement occurred only when decoding with a lexicon and LM (003M023F). Worse WERs with detoriations of approx. $1.6\% - 2.3\%$ were more likely to occur for systems $ASR_{02}(PDET_{02})$ and $ASR_{02}(PDET_{012})$ which were based on ≈ 65 more character tokens in comparison to the baseline systems. Overall, the results indicate that the prominence-aware ASR systems have comparable performance to the baseline systems.

[15]Note that these WERs are also similar to conversation-dependent mean WERs of $22.79\% - 29.81\%$ as described in Sec. 4.2.4

Table 4.11: WERs [%] of two conversations (003M023F/004M024F) for 1) baseline experiments (cf. Sec. 4.2.4), 2) ASR experiments based on prominence annotations from the prominence detector $PDET_{02}$ and 3) ASR experiments based on prominence annotations from prominence detector $PDET_{012}$.

Type	Lexfree 003M023F/004M024F	Lex 003M023F/004M024F	3-gram 003M023F/004M024F
Baseline	**26.04** / **31.25**	**21.78** / **27.52**	18.57 / **23.71**
$ASR_0(PDET_{02})$	26.54 / 32.32	22.31 / 28.64	18.58 / 24.50
$ASR_2(PDET_{02})$	26.27 / 32.34	22.24 / 28.31	18.50 / 24.32
$ASR_{02}(PDET_{02})$	26.66 / 33.33	23.92 / 29.84	18.95 / 25.61
$ASR_0(PDET_{012})$	26.88 / 32.48	21.86 / 29.03	**18.23** / 24.88
$ASR_2(PDET_{012})$	27.21 / 32.26	22.20 / 28.16	18.75 / 24.42
$ASR_{02}(PDET_{012})$	26.87 / 32.85	22.67 / 29.16	18.68 / 24.80

4.5.4 Discussion and conclusion

In this section, we demonstrated our efforts to develop a prominence-aware ASR system by integrating prosodic prominence information. We began by training prominence detectors with human-annotated data (cf. Fig. 4.24) to classify word prominence levels, which then formed the basis for developing ASR models that incorporate this prosodic information.

The prominence detection results show similar trends as the prominence classification results described in Sec. 3.3, even though the results are not directly comparable because of differences in the utilized data (cf. Sec. 3.3.5) and the evaluation methods. However, the main advantage of our prominence detection approach presented here is that no forced alignments are necessary, as word boundaries are detected automatically. Our results show that PERs and accuracies were best in case of the $PDET_{02}$ model (with PERs of $24.83\% \pm 1.79\%$ and accuracies of $89.72\% \pm 3.26\%$ for the word-aligned data) indicating that promising detection results can be achieved for both prominence levels. However, the $PDET_{012}$ model achieved worse PERs of $36.54\% \pm 0.92\%$ and accuracies of $69.45\% \pm 2.11\%$ for the word-aligned data which also illustrates the issues seen in the inter-annotator agreements which had Cohen's kappa of 0.72 and 0.57 with respect to *PL1* (cf. Sec. 3.3.1). In comparison, Heckmann et al. (2014) found that despite using different HMM-based alignment strategies for prominence detection, the unweighted accuracies for distinguishing prominent from non-prominent words with prosodic features were approx. $80\% - 82\%$, which is consistent with our findings. Whereas our prominence detector aligns speech directly to a sequence of prominence levels, the methods described in Heckmann et al. (2014) rely on forced alignments that require text transcriptions as input in order to train prominence classification models. This also implies that their evaluation assumes that all words can be consistently aligned with the human annotations. In conclusion, prominence detection on conversational speech with wav2vec2 works well even without requiring forced alignments to detect phone or word boundaries. More precisely, our results indicate that fine-tuned speech representation models automatically extract representations that capture prosodic information. This prosodic information can then be used for downstream ASR tasks.

The incorporation of the detected prominence information did not lead to improvements in ASR performance, but enabled the training of prominence-aware

ASR systems which also output prosodic information. Independent of the decoding strategy (without/with lexicon/LM), the additional word-level prominence information mapped onto the character-level in general led to consistent results when comparing the WERs to the baseline. However, slightly worse results were achieved for those ASR models where more character tokens were involved.

To conclude, our study demonstrates that prominence detection in conversational speech using wav2vec2 is feasible without relying on forced alignments, as the model effectively extracts prosodic information automatically. When using wav2vec2 for transcribing words and prominence levels simultaneously, the explicit information about prominence levels did neither enhance nor deteriorate ASR performance, while additionally providing labels for prominence levels. To the best of our knowledge, this kind of prosody-enhanced ASR transcript is a novel contribution to the field, with high relevance to both speech science and speech technology. In future work, it would be interesting to examine the models including only the strong/empathic prominence information (i.e., *PL2*), as this could be useful for applications such as automatic annotation for linguistic research, and prosody-informed natural language understanding (NLU) components for dialogue systems or comprehension aids.

Chapter 5

General discussion and conclusion

This discussion addresses the two research aims and their corresponding research questions formulated in the introduction (cf. Sec. 1.1) and summarizes the main contributions and findings from my thesis.

5.1 Analysis of acoustic representations and models for conversational speech with explainable machine learning methods

For the first aim of this thesis, we addressed several research questions related to the analysis of acoustic representations and models for conversational speech with explainable machine learning methods (cf. Sec. 1.1.1). At first, we focused on analyzing the main acoustic cues of prosodic prominence for conversational speech (**RQ1**; cf. Sec. 5.1.1). Second, we examined whether WERs of conversational speech are affected by utterance-level features (**RQ2**; cf. Sec. 5.1.2). Third, we explored what shared discrete speech representations encode with respect to language varieties, speaking styles, and speakers (**RQ3**; cf. Sec. 5.1.3). Finally, we investigated whether the fine-tuning of self-supervised speech representations implicitly encodes prosody (**RQ4**; cf. Sec. 5.1.4).

5.1.1 Main acoustic cues for prosodic prominence

In Sec. 3.3 we conducted an analysis of prominence classification tools by training explainable random forest models. The findings highlighted the critical role of durational characteristics for prominence classification. Moreover, we proposed entropy-based features that allow the models to maintain the same level of accuracy without relying on the calculation of more sophisticated durational features, thereby simplifying the classification process while preserving performance.

The role of durational features: The study presented in Sec. 3.3 demonstrated that durational features are necessary for word-level prominence classification in

conversational speech. The comparison between a basic prosodic feature set and a subset without any durational features showed that cross-validation recalls of non-prominent words improved in case of the basic feature set ($81\% \pm 3\% > 72\% \pm 4\%$). Furthermore, the results also indicates that word duration has by far the highest importance with respect to the impurity-based feature ranking of the random forest. Previous studies on prosodic prominence pointed towards different trends. Cole et al. (2010) found that listeners perceive words as prominent when corresponding stressed vowels had longer durations. In contrast, research by Niebuhr and Winkler (2017) demonstrated that, when manipulating F0 and duration, F0 serves as a stronger indicator of perceived prominence in German than duration. Then again, the analysis of variable importance in Baumann and Winter (2018) showed that RMS is the most important feature among all continuous-valued acoustic variables. In summary, our results emphasize that especially word duration plays an important role in classifying word-level prosodic prominence.

The role of entropy-based features: We show that novel entropy-based features (i.e., pseudo-entropies) based on F0 and RMS contours encode necessary durational information in order to classify word-level prominence in conversational speech (cf. Sec. 3.3, Sec. 4.2.3 or Appendix A). Our cross-validation results demonstrate that classification performance remained consistent across three feature sets: a basic prosodic feature set (96 features), a subset in which entropy-based features replaced durational features (88 features), and an expanded prosodic feature set that included additional entropy-based features (100 features). To the best of our knowledge, no prior studies have investigated comparable entropy-based F0 and RMS features for prominence classification. To conclude, classifying prominence levels in conversational speech using novel pseudo-entropies show that the calculation of phone-based durational features can be omitted, as the necessary durational information is encoded within these entropy measures.

5.1.2 Effects on WERs in conversational speech

In Sec. 4.2 we compared the ASR performances in conversational speech with respect to the ASR architectures Kaldi, wav2vec2 and Whisper. To gain a deeper understanding of the factors influencing their performance, we investigated the impact on utterance-level WERs with respect to utterance length, prosodic, pronunciation and perplexity features.

The role of utterance length and articulation rate: We identified that the utterance length in number of word tokens and the articulation rate significantly affect utterance-level WERs in the ASR systems Whisper and Kaldi. When comparing the overall distributions of the mean WERs at utterance level with respect to the number of word tokens, we observed that Whisper achieved worse WERs for short utterances (mean WER of 43.7%) and better WERs for long utterances (mean WER of 37.5%). In contrast, this direction was reversed for all other ASR systems (Kaldi: $30.2\% < 43.5\%$; wav2vec2 without lexicon/LM: $22.4\% < 30.2\%$; wav2vec2 with lexicon/LM: $16.8\% < 23\%$). With respect to articulation rates, the mean WERs on utterance level for single-word utterances were usually best for lower articulation rates, except for Whisper. Notably, in case of the zero-shot

Whisper system, there was a *sweet spot* with good WERs for articulation rates between $10\,\mathrm{s}^{-1} - 12\,\mathrm{s}^{-1}$. For the other architectures, mean WERs were generally better in case of shorter utterances with articulation rates $< 12\,\mathrm{s}^{-1}$ (with slight exceptions in case of Kaldi for three or four word tokens). Overall, the systems having linguistic information in form of an LM, a pronunciation lexicon (Kaldi) or a simple word-level lexicon (wav2vec2 with lexicon/LM) are most robust against high articulation rates. Hirschberg et al. (2004) found that HMM-based systems performed worse on longer turns than shorter ones. However, our results indicate that modern hybrid DNN-HMM (Kaldi) and transformer-based (wav2vec2) models are more robust to utterance length with respect to the number of word tokens. ASR systems based on Kaldi and wav2vec2 achieved nearly constant mean WERs for utterances of $2 \ldots 15$ word tokens (Kaldi: $\approx 40\%$; wav2vec2 without lexicon/LM: $\approx 30\%$; wav2vec2 with lexicon/LM: $\approx 25\%$). Furthermore, Wei et al. (2022) demonstrated that a transformer-based conversational ASR system benefited from more contextual information which is in line with our findings. We observe that increased context at the utterance level improved recognition results for the zero-shot Whisper system. Regarding articulation rate, Goldwater et al. (2008) reported little effect on WERs for words close to the average speech rate but more errors for extreme values. They also found the fewest errors for words longer than average. Their observations on HMM-based systems align with our findings, particularly for Whisper, where we identified a *highly significant* interaction between the number of word tokens and the articulation rate. Specifically, for two and three word tokens, Whisper exhibited a *sweet spot* at utterance-level WERs of 50% for articulation rates between approx. $10\,\mathrm{s}^{-1} - 12\,\mathrm{s}^{-1}$. In summary, our analysis revealed that utterance length and high articulation rate have a noticeable effect on ASR performance. The zero-shot Whisper system performed worse than Kaldi and wav2vec2, especially in case of short utterances.

The role of entropy-based F0 and RMS features: In Sec. 3.3 we introduced entropy-based features (Linke, Kubin, & Schuppler, 2023) that encode contour variation of F0 and RMS (i.e., pseudo-entropies). Our statistical analysis revealed a strong negative correlation between utterance-level WERs and the pseudo-entropies of F0 for short utterances (one to four word tokens), with Whisper exhibiting the strongest effect. Specifically, utterances with fewer word tokens and more uniform F0 contours (lower pseudo-entropies) tended to have lower WERs. This trend was most pronounced for single-word utterances and for Whisper in particular, where a more uniform F0 contour corresponded to notably lower WERs. In contrast, wav2vec2 and Kaldi demonstrated greater independence from the F0 contour, especially for utterances containing two to four word tokens. Furthermore, we discovered a significant interaction between pseudo-entropies of F0 and articulation rates, revealing that the best WERs were achieved when flat F0 contours were combined with slow speech. To the best of our knowledge, the influence of F0 and RMS contour entropies on ASR performance has not been investigated in previous studies. Goldwater et al. (2008) analyzed errors of an HMM-based ASR system and discovered that *"more extreme values"* of pitch mean and range were associated with higher WERs. Their findings align with our results, which demonstrated that utterances with less uniform F0 contours were associated with higher WERs. In summary, all ASR architectures exhibited sensitivity to F0 variation, particularly in

case of single-word utterances. However, for utterances containing two to four word tokens, the performance of ASR systems trained or fine-tuned on domain-specific data was unaffected by F0 variation.

The role of pronunciation variation: Since the speakers in the conversational speech material spoke a regional variety of the German language, our WER analysis investigated the impact of pronunciation variations on ASR performance. We measured the mean Levenshtein distance between the realized pronunciation of a word to its canonical pronunciation in the lexicon for standard Austrian German. This feature exhibited a strong correlation with WERs of the zero-shot Whisper system, which can be attributed to its lack of fine-tuning on in-domain data and the absence of a pronunciation lexicon. For all ASR systems, utterance-level WERs were best when pronunciations were closer to standard Austrian German. Whisper had higher WERs for both short and long utterances when pronunciations were further away from the standard. In contrast, for shorter utterances the WERs of the other ASR systems were less affected by pronunciation variation. Interestingly, wav2vec2 without lexicon/LM performed better for utterances with three to four words and mean Levenshtein distances between approx. $1-2$, possibly due to its transformer encoder which benefits from larger context. As expected, wav2vec2 with lexicon/LM was slightly more robust against pronunciation variation than without a lexicon/LM. Prior research has highlighted the influence of phonetic neighborhood density on the performance of HMM-based ASR systems. Goldwater et al. (2008) discovered that high phonetic neighborhood density can lead to increased recognition difficulties. However, our study takes a different approach by focusing on conversational speech from a low-resourced language variety. We aimed to assess the extent to which the pronunciation of utterances deviated from standard norms, a factor that has not been extensively explored in earlier studies. This perspective is especially important considering that most German speech models are trained on non-Austrian data, often using prepared speech. We hypothesized that utterances with pronunciations closer to the standard would be recognized with greater accuracy. Our results confirm this hypothesis, particularly in light of Whisper's reduced performance on utterances spoken with a pronunciation that diverged more from standard Austrian German. Furthermore, we observe that incorporating a knowledge-based lexicon on top of a transformer-based system is beneficial.

The role of perplexity: We analyzed the effect of utterance-level perplexities (calculated with a four-gram LM trained on a subset of 5 M German sentences from Wikipedia and the European parliament) on WERs but found weaker effects in comparison to all other features referring to utterance length, prosody and pronunciation. Nevertheless, our analysis indicated that WERs tended to be lower for short utterances with lower perplexities across all ASR architectures. Moreover, our analysis revealed a slight deterioration in WERs with increasing perplexity for utterances containing two to four word tokens. Unsurprisingly, we also discovered that 'not surprising' word sequences spoken closer to the standard pronunciation led to better WERs in comparison to utterances with 'more surprising' word sequences spoken further away from the standard. In their study, Goldwater et al. (2008) also found an almost linear relationship between trigram-log-probabilities and WERs. Our findings are consistent with this, as we found that lower perplexities (equivalent

to higher log-probabilities) generally resulted in better WERs. However, while we did observe an effect of LM probabilities on WERs, it was weaker compared to the influence of other utterance-level features such as utterance length, articulation rate, pseudo-entropies of F0, and pronunciation Levenshtein distance. This trend was consistent across all ASR architectures investigated. We recognize that the simple n-gram language modeling approach used to calculate the perplexity feature has limitations and may not be the most effective method for estimating language model probabilities, particularly when it comes to capturing the nuances of the conversational speaking style.

5.1.3 Towards the encodings of shared discrete speech representations

In Sec. 4.4 we analyzed self-supervised pre-trained shared discrete speech representations with respect to different speech corpora. Driven by the success of fine-tuning self-supervised representations for ASR in low-resource settings (Baevski, Zhou, et al., 2020) and their cross-language sharing capabilities (Conneau et al., 2021), our research aimed to investigate their broader encoding potential. Our investigations revealed that these representations encode not only different languages but also language varieties and speaking styles. Additionally, we demonstrated that the same speakers exhibit different behaviors across speaking styles, highlighting the nuanced capabilities of these representations in capturing diverse speech characteristics.

The role of languages, varieties and speaking styles: We analyzed shared discrete speech representations from speakers of different corpora with respect to languages, varieties and speaking styles. The clustering experiment with XLSR codebook entries from the different data sets showed that, in addition to languages, read and spontaneous speaking styles are effectively differentiated in this feature space. In summary, we found that read speech allows for effective differentiation between German and Austrian German varieties, independent of language. However, spontaneous speech proved more challenging to classify by variety, with only the German face-to-face conversations and Hungarian task-oriented/task-restricted conversations showing good clustering performance (F1-scores of 77% and 78%). We attribute this variety-independence in spontaneous German speech to the highly varied usage of speech representations. Our results align with previous research, such as Conneau et al. (2021), who successfully clustered discrete speech representations from multilingual pre-trained wav2vec2 models to group related languages using similar methods. Similarly, Sato and Heffernan (2020) successfully clustered Japanese dialects using sentence vector representations based on character-based metrics, identifying three distinct clusters: Tohoku dialect, Tokyo dialect, and a combination of three Western dialects (Kansai, Chugoku, and Kyushu). In conclusion, our study analyzed the speaker-dependent usage of discrete shared speech representations and we found that they effectively differentiate languages (German vs. Hungarian), German and Hungarian speaking styles (read speech vs. spontaneous speech) and German varieties in read speech (Austrian German vs. Northern German). The clustering of German varieties in spontaneous speech proved to be more challenging and indicated a degree of variety-independence.

The role of same speakers in different styles: Another focus of our analysis was on how the speech representations of the same speakers behave and whether they explain different degrees of spontaneity. We found that Austrian German speakers differ the most between different styles since mean distance of same Austrian German speakers was high (2.3). In contrast, mean distance of same Hungarian speakers was smaller (1.3). Furthermore, we found that Austrian German speakers also differ more from themselves within different styles, indicating speaker identity independence of the speech representations. Overall, our results indicate that speech representations vary the most among Austrian German speakers. Also Asami et al. (2014) found that GMM supervectors based on utterances can discriminate read and spontaneous speech with less speaker-dependency. Simultaneously, the authors state that clustering spontaneous utterances is more difficult than read utterances. To conclude, the results suggest that distance calculation based on shared quantized latent speech representations is also meaningful on a much finer granularity level (i.e., per speaker per speaking style) than it was introduced in Conneau et al. (2021) for languages. This may open new perspectives in speech data selection both for supervised and self-supervised learning, as speech sections matching the desired development set (or speaking style) could be collected at a relative low cost, requiring only a pre-trained wav2vec2 model but without the need of any additional information beyond the waveform. Furthermore, it may be worth exploring meaningful acoustic correlates that could shed more light on the nature of elusive self-supervised speech representations. We plan to extend our investigations in these directions in the future.

5.1.4 Towards prosody of fine-tuned speech representations

In Sec. 4.5, we introduced prominence-aware ASR systems designed to incorporate prosodic information into the speech recognition process. By integrating automatically detected prominence levels into transformer-based ASR systems, the results suggest that these systems process and interpret prosodic conversational speech patterns. While the results did not show an improvement in overall ASR performance, we have demonstrated that it is possible to train prominence-aware ASR systems without compromising the quality of the output. Hence, our ASR results indicated that prominence-aware ASR systems remained consistent with the baseline, regardless of the decoding method employed (with/without lexicon/LM). This consistency suggests that the integration of prominence information does not adversely affect the core functionality of fine-tuned transformer-based ASR systems. Furthermore, the findings indicate that the self-attention mechanisms and context networks employed by transformer-based ASR systems may be capable of implicitly modeling and processing prosodic patterns. To the best of our knowledge, this research represents a novel contribution to the field which bridges the fields of speech science and speech technology. Thus, the integration of prominence information into ASR systems opens up exciting future directions, enabling researchers to explore the potential benefits of this approach in linguistic research, dialogue systems, and comprehension aids.

5.1.5 Future work

The insights gained from this thesis open future directions for several areas of research in speech science and speech technology. Future work should focus on further refining the entropy-based prosodic features and investigating their potential for improving ASR performance, especially for conversational speech with high articulation rates and regional pronunciation variations.

The clustering analysis of shared discrete speech representations should be extended to a wider range of languages, varieties, and speaking styles. Our clustering approach can also be applied to investigate other speech phenomena such as prosodic prominence or pitch patterns. Recent work on Hungarian speech recognition has already employed our analysis approach to visualize the speech data composition and differentiate between spontaneous and non-spontaneous speech styles (Mihajlik et al., 2023). Furthermore, we are collaborating with experts from language documentation (Alexander Zahrer, University of Münster) with whom we analyze speech representations of speakers from Papua, a region known for its linguistic diversity. This joint effort aims to uncover speaker-, language- and variety-dependent patterns that could potentially enhance annotation strategies and guide future fieldwork in this area.

Our prominence-aware ASR approach shows the feasibility of integrating prosodic information into ASR systems. Future work should explore more sophisticated methods for incorporating this information and assess its potential impact on linguistic research and dialogue systems. Our studies on prominence detection are ongoing and currently applied to charisma research, i.e., to the analysis on how prosodic prominence correlates with perceived speaker charisma (in collaboration with Oliver Niebuhr, University of Southern Denmark). Finally, to better comprehend the underlying nature of (fine-tuned) self-supervised speech representations or transformer-based ASR architectures in general, future work should explore their relationship with other prosodic phenomena, potentially uncovering insights into their internal mechanisms and representations of prosody, following the research line examplified by Shim et al. (2022), ten Bosch et al. (2023) or de Heer Kloots and Zuidema (2024).

5.2 Evaluation of acoustic representations and models for conversational speech with standard performance measurements

The second aim of this thesis addressed several research questions related to the evaluation of acoustic representations and models for conversational speech using standard performance measurements (cf. Sec. 1.1.2). We began by investigating whether word-level prominence classification results with prosodic features or word-level prominence detection results with fine-tuned speech representations align with inner-annotator agreements (**RQ5**; cf. Sec. 5.2.1). Subsequently, we explored how low-resourced HMM-based ASR systems compare to low-resourced or data-driven transformer-based ASR systems in terms of their effectiveness for recognizing Austrian German conversational speech (**RQ6**; cf. Sec. 5.2.2).

5.2.1 Automatic annotation of prosodic prominence for conversational speech

In Sec. 3.3 we introduced a prominence classification tool for conversational speech, which utilized prosodic features as its foundation. Building upon this work, we advanced our approach in Sec. 4.5.2 by developing a more sophisticated prominence detection tool that operates directly on raw audio data. Our results indicate that the performance of both of these annotation tools align with human inter-rater agreements.

Automatic Prominence Classification Our prominence classification results using prosodic features and random forest models are consistent with human inter-rater agreements. The model's best performance for three prominence levels (cross-validation accuracy: $63\% \pm 7\%$) can be attributed to the agreements between non-prominent and weakly-prominent words (0.72) and between weakly-prominent and highly-prominent words (0.57). Similarly, the model's best results for the two classes (cross-validation accuracy: $88\% \pm 5\%$) can be explained by the high agreement between non-prominent and highly-prominent words (0.92). The overall Cohen's kappa of these human inter-rater agreement was 0.72 which is similar to those reported in other studies, such as 0.53 by Tamburini and Wagner (2007) and 0.84 by Baumann and Winter (2018). The strong performance of our models with two classes suggests that prominence classification can be used as an automatic annotation tool. Additionally, our results suggest that carefully designed feature sets can eliminate the need for more complex durational features relying on forced alignments. However, our results also highlight that pitch detection can be problematic in conversational speech with shorter utterances (cf. Sec. 3.3.5). This emphasizes the importance of focusing on features that can be extracted and calculated robustly, to subsequently also ensure robust prominence classification in various speech contexts.

Automatic Prominence Detection The prominence detectors based on wav2vec2 show results that follow similar trends to the prominence classification results, although the results are not directly comparable (cf. Sec. 4.5.2). Notably, these prominence detection results also align with human inter-rater agreements. The prominence detector for three classes achieved prominence detection error rates of $36.54\% \pm 0.92\%$ and cross-validation accuracies of $69.45\% \pm 2.11\%$ for the word-aligned data. These results illustrate the issues observed in the agreements between non-prominent and weakly-prominent words (0.72) and between weakly-prominent and highly-prominent words (0.57). In contrast, our best results of the two-level prominence detector showed prominence detection error rates of $24.83\% \pm 1.79\%$ and cross-validation accuracies of $89.72\% \pm 3.26\%$ for the word-aligned data, which can be attributed to the high agreement between non-prominent and highly-prominent words (0.92). Heckmann et al. (2014) found that despite using different HMM-based alignment strategies for prominence detection, the unweighted accuracies for distinguishing prominent from non-prominent words with prosodic features were approx. $80\% - 82\%$, consistent with our findings. However, our prominence detector directly aligns speech to a sequence of prominence levels, whereas the methods described by Heckmann et al. (2014) rely on forced alignments that require text transcriptions as input to train prominence classification models. This implies that their evaluation

assumes that all words can be consistently aligned with human annotations. In conclusion, prominence detection on conversational speech using wav2vec2 performs well without requiring forced alignments to detect phone or word boundaries.

5.2.2 A comparison of ASR architectures for conversational speech

This thesis investigated the performance of different automatic speech recognition architectures when processing conversational speech in Austrian German. Our research revealed that a low-resourced language processing assumption is supported for Austrian German conversational speech (cf. Sec. 4.3). Additionally, we examined the robustness of different ASR systems with distinct characteristics by comparing their performance on read vs. conversational Austrian German (cf. Sec. 4.2.4).

How much training data does ASR for conversational speech require? In general, this thesis provides conversation-dependent WERs with respect to three different ASR architectures (Kaldi, wav2vec2 and Whisper) and four different training approaches. When training HMM-based systems solely with GRASS CS and a cross-entropy loss (Kaldi), the best WERs were $51.87\% \pm 4.83\%$ or $56.19\% \pm 5.4\%$[1] (cf. Sec. 2.3.2 and Sec. 4.3.3). Then again, when training HMM-based systems with the LF-MMI criterion (Kaldi), we achieved better WERs of $42.86\% \pm 4.78\%$ (cf. Sec. 4.2.4). When using the same GRASS CS speech corpus for both pre-training and fine-tuning of the wav2vec2 architecture, we observed differences in performance with respect to the target sets: For the phone-based model, WERs were $57.28 \pm 6.46\%$ while the character-based model yielded WERs of $62.54\% \pm 6.36\%$[1] (cf. Sec. 4.3). In contrast, fine-tuning wav2vec2 models pre-trained on 56 000 h of multilingual speech data resulted in WERs of $25.06\% \pm 4.42\%$[1] or $22.79\% \pm 4.02\%$ (cf. Sec. 4.3 and Sec. 4.2.4). Interestingly, a zero-shot Whisper system pre-trained on enormous amount of multilingual (out-of-domain) speech data achieved WERs of only $41.78\% \pm 8.23\%$ (cf. Sec. 4.2.4). Hence, independent of the utilized data and independent of the ASR architecture, all results showed different means and especially high standard deviations for the conversation-dependent WERs. Likewise, these results demonstrate that a zero-shot ASR system (Whisper) which was trained on enormous amounts of multilingual (out-of-domain) speech data (680 000 h) and low-resourced ASR systems, which were trained entirely on (in-domain) speech data (a share of approx. $\frac{13.5\,\text{h}}{680\,000\,\text{h}} \hat{=} 0.002\%$), both achieve poor performance for Austrian German conversational speech (means of conversation-dependent WERs were $> 40\%$). Simultaneously, fine-tuning the wav2vec2 architecture (pre-trained on 56 000 h of multilingual speech data) with (in-domain) speech data (a share of approx. $\frac{13.5\,\text{h}}{56\,000\,\text{h}} \hat{=} 0.025\%$) improved the mean of the conversation-dependent WERs by approx. 20% but the standard deviation was still high at approx. 4%.

[1] Note that experiments described in Sec. 4.3 were based on the initial version of GRASS CS, in contrast to the other experiments, which were based on an updated version of GRASS with partial corrections of human annotations. However, this did not affect the data's comparability to other experiments, as the results align with all other findings. For instance, the Kaldi results presented in Sec. 4.3.3.1 were based on a lexicon with many pronunciation variants which yielded to similar results as the Kaldi results presented in Sec. 2.3.2 which were also based on a similar lexicon (i.e., *allPVs*).

In conclusion, our findings show that a) performing cross-validation by testing each conversation individually points out conversational speech complexity (i.e., large standard deviations for all ASR systems), b) fine-tuning a data-driven pre-trained cross-lingual speech representation model is effective for Austrian conversational speech (cf. fine-tuned wav2vec2 results), c) fine-tuning a low-resourced speech representation model pre-trained only on Austrian conversational speech is not effective for Austrian conversational speech and d) decoding with a zero-shot Whisper model is not effective for Austrian conversational speech. These findings from a) to d) together support a low-resourced language processing assumption for the Austrian German conversational speaking style.

How robust are ASR systems? In our comparison of the four different ASR systems (Kaldi, wav2vec2 and Whisper), Whisper was the only ASR system that was not fine-tuned, making it an example of a system not having any in-domain data. In contrast, the other systems were informed with in-domain speech data, with wav2vec2 employing two distinct decoding strategies: w2v (without lexicon/LM) and w2vLM (with lexicon/LM). Our analysis revealed that all ASR systems performed well on Austrian read speech, with speaker-dependent WERs ranging from 1.01% to 11.8%. Specifically, we observed mean WERs of 11.8% (Whisper), 3.62% (Kaldi), 1.81% (w2v), and 1.01% (w2vLM). However, in Austrian conversational speech the systems exhibited varying performance across different conversations, with mean WERs of 41.78% (Whisper), 42.86% (Kaldi), 29.81% (w2v), and 22.79% (w2vLM). A closer examination of individual conversations revealed high Pearson correlation coefficients between conversation-dependent WERs, with values $> 60\%$ for Whisper compared to all other systems, and $> 87\%$ for Kaldi versus wav2vec2 comparisons. These results indicate that while the four ASR systems achieve state-of-the-art performance for non-spontaneous speech material, they lack robustness in recognizing casual conversational speech. The robustness of ASR systems across diverse speech contexts remains a crucial area of investigation. Radford et al. (2023) claimed that supervised speech recognition models trained exclusively on English Librispeech (Panayotov et al., 2015) exhibit very different robustness properties. They supported this claim by demonstrating Whisper's superior performance over previous ASR benchmarks on various English datasets, including Common Voice (Ardila et al., 2020) and Switchboard (J. J. Godfrey et al., 1992). Furthermore, they suggested that Whisper's error patterns potentially align with human annotation behavior, at least for English datasets, based on comparisons with 95% confidence intervals of human errors. However, our findings present a more nuanced perspective with respect to robustness. In our study, the Whisper system (large-v2), despite being trained on substantial German speech data, exhibited considerable variability in performance. We observed large variations in both speaker-dependent and conversation-dependent WERs, with absolute differences between means and standard deviations of approx. 30% and 5.46%. This variability is particularly noteworthy given that Whisper had previously achieved impressive results on German speech data, outperforming its English speech recognition capabilities (e.g., WERs of $5.5\% < 6.2\%$ for Multilingual Librispeech and $6.4\% < 9.5\%$ for Common Voice 9). Szymański et al. (2020) raised concerns about low WERs on benchmark data sets like, e.g., Switchboard (J. J. Godfrey et al., 1992) or Callhome. Their research on internal multi-domain benchmarks revealed considerably higher WERs, ranging from 13.73% (for an insurance domain)

to 22.16% (for booking and wireless telecommunication calls). These findings align with our observations, further emphasizing the persistent challenge of robustness in state-of-the-art ASR systems when confronted with diverse, unseen speakers or conversation types. In conclusion, our research underscores the ongoing robustness issues in ASR systems which have not been fully resolved. This is one of the main reasons why we were driven to explore the causes of the challenges in recognizing conversational speech across different ASR architectures.

5.2.3 Future work

The insights gained from this thesis open future directions for several areas of research in speech technology. First, future work can explore more sophisticated methods for the prominence-aware ASR systems. This should potentially improve performance and robustness, especially for casual conversational speech which often contains many short utterances. Second, future research should investigate the causes of the robustness issues observed in ASR systems when recognizing conversational speech. For instance, future work can extend our evaluation to other common spontaneous speech corpora by examining the variability of conversation-dependent WERs or comparing utterance-level WERs. This comprehensive approach would provide valuable insights into the factors affecting ASR performance across different conversational contexts and speaking styles. Finally, our work suggests that future work should focus on developing ASR architectures and training strategies specifically tailored for low-resourced languages in casual conversational speaking styles. Our study opens up several avenues for future research. First, future work should explore all possible combinations of system configurations. For example, researchers could investigate how wav2vec2 fine-tuned on a corpus of canonical German read speech performs in the conversational speech setting. Furthermore, given the peculiar performance of Whisper, an analysis of Whisper fine-tuned on GRASS is recommended. This would provide insights into the performance of a transformer-based ASR system without explicit linguistic knowledge that has been fine-tuned on data from the target language and style. Second, future experiments should extend beyond one corpus to enhance the generalizability of the findings. We anticipate that our analysis will draw attention to phenomena relevant for other conversational speech corpora and motivate researchers to work on conversational speech from various languages and dialects. This expansion of scope will contribute to a more comprehensive understanding of ASR systems in diverse linguistic contexts.

5.3 Conclusion

This thesis set out to analyze and evaluate acoustic representations and models for conversational speech for two tasks: prosodic prominence classification and ASR. My research was guided by two main aims: 1) to analyze acoustic representations and models using explainable machine learning methods, and 2) to evaluate these representations and models using standard performance measurements. The key contributions of my thesis are:

1. The most important cues to prosodic prominence in conversational speech are related to word duration and speech rate, and not to pitch (as has earlier been

shown for controlled experiments; cf. Arnold et al. (2013) or Niebuhr and Winkler (2017)). Entropy-based prosodic features encode these durational aspects along with F0 and RMS. When these are used for prominence classification, human-level performance was achieved.

2. A comprehensive analysis of the factors affecting ASR performance showed that HMM-based systems perform better than transformer-based systems for short utterances of large pronunciation variation, whereas transformer-based systems can deal better with long utterances independent of speech rate and F0 and RMS variation.

3. When clustering shared discrete speech representations, we can effectively differentiate not only languages (as shown before; cf. Conneau et al. (2021)), but also varieties, speaking styles and individual speakers performing different speaking styles.

4. Self-supervised representations encode information relevant to prosodic prominence. This thesis presents the first transformer-based prominence aware ASR system.

This thesis contributes to understanding the complexities of conversational speech processing and recognition, with implications for improving ASR systems, particularly for low-resourced languages and conversational speaking styles. My work may hopefully open new avenues for incorporating prosodic information into speech technology, not only the novel acoustic prosodic features developed, but also the analysis methods. As conversational AI continues to evolve, the insights gained from this research will hopefully inform the development of future robust speech recognition systems for natural, spontaneous dialogues between humans and machines.

Appendices

Appendix A: Comparison of entropy-based features

This Appendix compares the *conventional entropy* (as described in the literature) with the pseudo-entropies used in Sec. 3.3, Sec. 4.2 and Sec. 4.5 of this thesis. At first, the calculations of the two entropies are compared. Second, the resulting measurements are described by comparing different F0 contours of female GRASS speakers[2]. Finally, correlation matrices with respect to the data used for the CS prominence classification experiment (cf. Sec. 3.3) are shown.

Conventional entropy: Entropy describes the received information of the realization x_i of a random variable X. In general, entropy describes the *degree of surprise* by allocating very unlikely events with more information and very likely events with no information. Hence, the resulting information measure

$$H(X) = -\sum_i p_i \cdot \log p_i, \tag{5.1}$$

depends on a probability distribution where $p_i = Pr\{X = x_i\}$ describes the probability of the random variable X taking the value x_i. The maximum is

$$\begin{aligned} H_{max} &= -\sum_{i=1}^{N} \frac{1}{N} \cdot \log \frac{1}{N} \\ &= -\left(\frac{1}{N} \log \frac{1}{N} + \cdots + \frac{1}{N} \log \frac{1}{N} \right) \\ &= -\left(\frac{1}{N} \cdot N \cdot \log \frac{1}{N} \right) \\ &= -(\log 1 - \log N) = \log N. \end{aligned} \tag{5.2}$$

The minimum can be derived for $p_1 = 1$ and $p_i = 0$ for all $i \in \{2, \ldots, N\}$:

$$H_{min} = -p_1 \cdot \log p_1 = 0. \tag{5.3}$$

In case of the *conventional* entropy, we estimated probability distributions by creating histograms for F0 contours (by binning the feature values between 80 Hz

[2]Note that in the description of those contours, conventional entropies are referred to as **HF0** while pseudo-entropies are referred to as **HPSF0**.

and 240 Hz with a width of 5 Hz) and RMS contours (by binning the feature values between 0 and 1 with a width of 0.05). In both cases the probabilities were normalized such that the $\sum_{i=1}^{N} p_i = 1$.

Pseudo-entropy: We defined the entropy-based features by utilizing again the entropy formula as described in Eq. 5.1). In contrast to estimating a probability distribution via binning, we can define the (pseudo-)probability distribution for a sequence of N (non-negative) feature values $f[i]$ by calculating

$$p_i = \frac{f[i]}{\sum_{n=1}^{N} f[n]}$$

These (pseudo-)probabilities also fulfill the total probability condition. The maximum relates to a constant contour (i.e., a uniform distribution) with $p_i = \frac{1}{N}$ which also results in $H_{max} = \log N$ (see Eq. 5.3). The theoretical minimum can be derived for a sequence of N (non-negative) feature values with

$$f[i] = \delta[i-1] = \begin{cases} 1, & i = 1 \\ 0, & i > 1 \end{cases} \tag{5.4}$$

which leads to the (pseudo-)probabilities

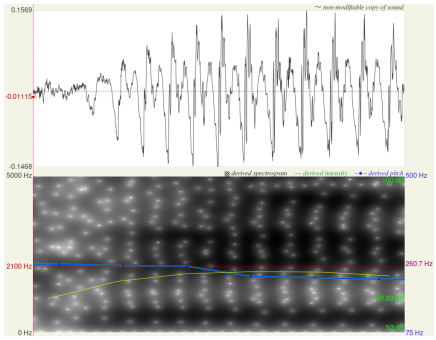
$$p_i = \begin{cases} 1, & i = 1 \\ 0, & i > 1. \end{cases}$$

These probabilities essentially describe the minimum for a sequence with $N = 1$ which also results in $H_{min} = 0$ (see Eq. 5.3). Nevertheless, for prosody experiments the theoretical minimum of the (pseudo-)entropy is not realistic because the sequence of feature values $f[i]$ represent F0/RMS contours which are generally not related to an impulse 5.4. In a more realistic scenario the minimum could refer to a power law of the form $f[i] = i^d$ which leads to (pseudo-)probabilites of the form

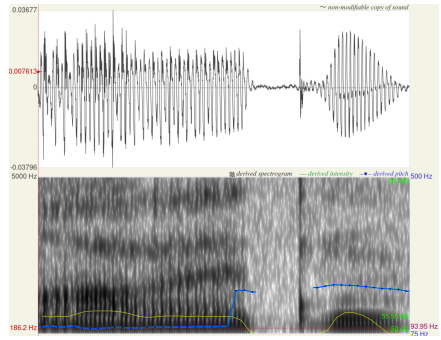
$$p_i = \frac{i^d}{\sum_{n=1}^{N} n^d}. \tag{5.5}$$

Note that p_i is non-negative since $i \geqslant 1$ and that p_i has its maximum at $i = N$. Furthermore, the order of the feature values does not influence the entropy calculation.

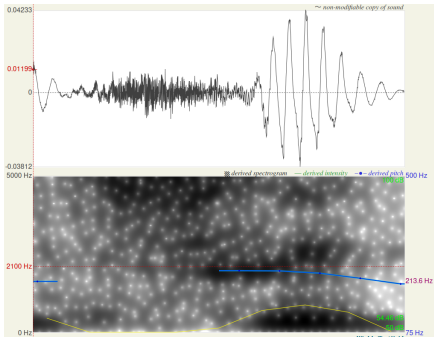
Entropy measurements for exemplary F0 contours: Fig. A.1 shows exemplary F0 contours with flat and medial peak pitch types of female GRASS speakers. For F0 contours with duration ≈ 0.1 s both conventional entropy values were 0. In contrast, pseudo-entropy values **HPSF0** were 1.386294 for the flat F0 contour and 0.693147 for the F0 contour annotated with a medial peak. In the first case, the higher value of **HPSF0** can be explained by the fact that more F0 values were detected. In case of words with longer durations of ≈ 0.5 s we observe a different trend. Here conventional entropies **HF0** had values of 1.539157 (flat) and 1.940843 (medial peak) indicating a higher degree of variation especially in the latter case. In contrast, corresponding pseudo-entropies **HPSF0** had values of 3.509627 (flat) and 3.123620 (medial peak). This reduction in **HPSF0** value for the latter case can be explained by the fact that fewer F0 values were detected.



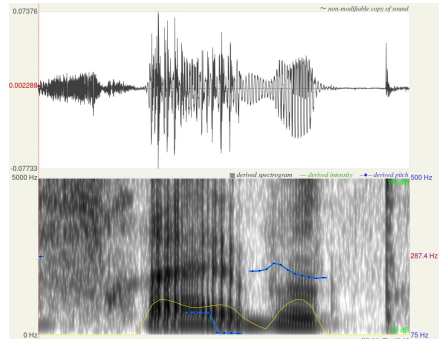
(a) Flat F0 contour with duration ≈ 0.1 s (German word "reden").



(b) Flat F0 contour with duration ≈ 0.5 s (German word "arbeiten").



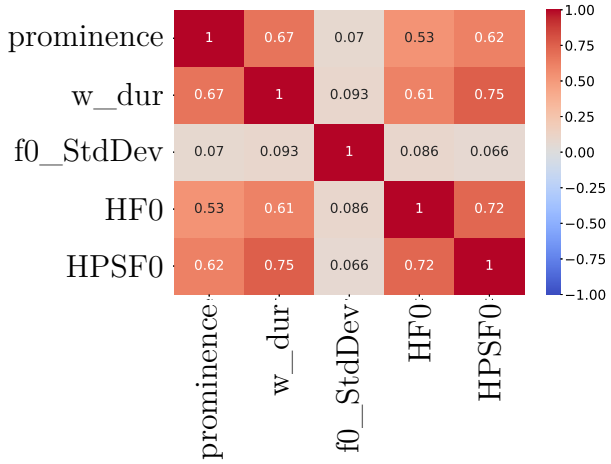
(c) F0 contour with duration ≈ 0.1 s and medial peak (German word "zur").



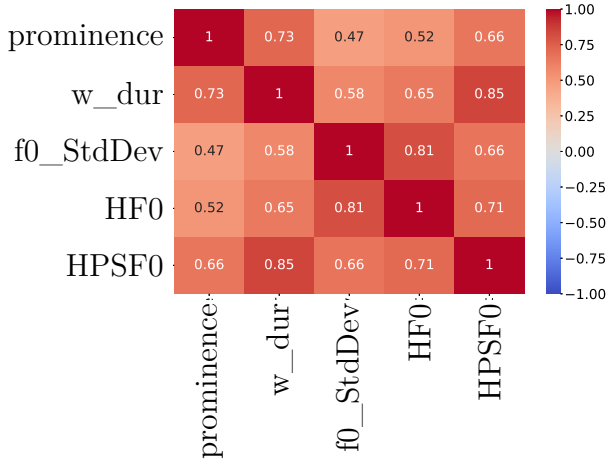
(d) F0 contour with duration ≈ 0.5 s and medial peak (German word "freibad").

Figure A.1: Exemplary F0 contours (blue) from female GRASS CS speakers. (a) shows a F0 contour annotated as a flat pitch type with duration ≈ 0.1 s and entropy values of 0 (**HF0**) and 1.386294 (**HPSF0**). (b) shows a F0 contour annotated as a flat pitch type with duration ≈ 0.5 s and entropy values of 1.539157 (**HF0**) and 3.509627 (**HPSF0**). (c) shows a F0 contour annotated as a medial-peak pitch type with duration ≈ 0.1 s and entropy values of 0 (**HF0**) and 0.693147 (**HPSF0**). (d) shows a F0 contour annotated as a medial-peak pitch type with duration ≈ 0.5 s and entropy values of 1.940843 (**HF0**) and 3.123620 (**HPSF0**).

Correlation matrices: Fig. A.2 shows correlation matrices with respect to Pearson and Spearman correlations between word prominence level (prominence), word duration (w_dur), the standard deviation of F0 (f0_StdDev), the conventional entropy (**HF0**) of F0 and the pseudo-entropy of F0 (**HPSF0**). Interestingly, with respect to Pearson, the feature f0_StdDev had no linear relationship with all other features. However, we observe that **HPSF0** correlates more strongly with word prominence ($0.62 > 0.53$) and w_dur ($0.75 > 0.61$) than **HF0**. The Spearman correlations showed similar trends, but the difference between **HF0** and **HPSF0** was even greater for w_dur ($0.85 > 0.65$). Furthermore, this time the Spearman correlation between f0_StdDev and all other features was larger, with f0_StdDev correlating most strongly with **HF0** (0.81).



(a) Pearson correlations



(b) Spearman correlations

Figure A.2: Correlation matrices showing the (a) Pearson and (b) Spearman correlations between word prominence level (prominence), word duration (w_dur), the standard deviation of F0 (f0_StdDev), the conventional entropy (**HF0**) and the pseudo-entropy (**HPSF0**).

Bibliography

Adda-Decker, M., & Lamel, L. (2018). Rethinking reduction: Discovering speech reductions across speaking styles and languages. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Interdisciplinary Perspectives on Conditions, Mechanisms, and Domains for Phonetic Variation* (pp. 101–128). Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110524178-004

Adda-Decker, M., Schuppler, B., Lamel, L., Morales-Cordovilla, J. A., & Adda, G. (2013). What we can learn from ASR errors about low-resourced languages: A case-study of Luxembourgish and Austrian. In *ERRARE Workshop - Ermenonville, Paris, France.*

Ananthakrishnan, S., & Narayanan, S. S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Trans. Audio Speech Lang Processing*, *16*(1), 216–228.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proc. of LREC* (pp. 4218–4222).

Arnold, D., Möbius, B., & Wagner, P. (2012). Comparing word and syllable prominence rated by naive listeners. In *Proc. of Interspeech* (pp. 1877–1880).

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PloS one*, *12*(4), e0174623.

Arnold, D., Wagner, P., & Baayen, R. H. (2013). Using generalized additive models and random forests to model prosodic prominence in German. In *Proc. of Interspeech* (pp. 272–276).

Asami, T., Masumura, R., Masataki, H., & Sakauchi, S. (2014). Read and spontaneous speech classification based on variance of GMM supervectors. In *Proc. of Interspeech* (pp. 2375–2379). doi: 10.21437/Interspeech.2014-516

Avanzi, M., Lacheret-Dujour, A., & Victorri, B. (2008). ANALOR. A tool for semi-automatic annotation of French prosodic structure. In *Proc. of Speech Prosody* (pp. 119–122).

Baevski, A., Auli, M., & Mohamed, A. (2020). Effectiveness of self-supervised pre-training for speech recognition. *ArXiv*, *abs/1911.03912*.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 12449–

12460). Retrieved from https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *Proc. of ICLR.* Retrieved from http://arxiv.org/abs/1409.0473

Bahl, L., Brown, P., de Souza, P., & Mercer, R. (1988). Speech recognition with continuous-parameter hidden Markov models. In *Proc. of ICASSP* (pp. 40–43). doi: 10.1109/ICASSP.1988.196504

Basson, W., & Davel, M. (2012). Comparing grapheme-based and phoneme-based speech recognition for Afrikaans. In *Pattern Recognition Association of South Africa (PRASA)* (pp. 144–148). doi: 10.13140/RG.2.1.1733.4487

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Baumann, S., Niebuhr, O., & Schroeter, B. (2016). Acoustic cues to perceived prominence levels – evidence from German spontaneous speech. In *Proc. of Speech Prosody.*

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgements. *J. Phon*, *70*, 20–38.

Beckman, M. E. (1986). Stress and non-stress accent. *Netherlands Phonetic Archive*, *7*.

Bishop, J. (2012). Information structural expectations in the perception of prosodic prominence. In G. Elordieta & P. Prieto (Eds.), *Prosody and Meaning* (p. 239–270). Berlin: Mouton de Gruyter.

Braunschweiler, N. (2003). ProsAlign - The Automatic Prosodic Aligner. In *Proc. of ICPhS* (pp. 3093–3096).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1 [Computer software manual]. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., ... Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In S. Renals & S. Bengio (Eds.), *Machine Learning for Multimodal Interaction* (pp. 28–39). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chen, G., Xu, H., Wu, M., Povey, D., & Khudanpur, S. (2015). Pronunciation and silence probability modeling for ASR. In *Proc. of Interspeech* (pp. 533–537). doi: 10.21437/Interspeech.2015-198

Chiba, Y., & Higashinaka, R. (2021). Dialogue situation recognition for everyday conversation using multimodal information. In *Proc. of Interspeech* (pp. 241–245). doi: 10.21437/Interspeech.2021-171

Christensen, R. H. B. (2023). ordinal—regression models for ordinal data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ordinal (R package version 2023.12-4.1)

Christodoulides, G., Avanzi, M., & Simon, A. C. (2017). Automatic labelling of prosodic prominence, phrasing and disfluencies in French speech by simulating the perception of naïve and expert listeners. In *Proc. of Interspeech* (pp. 3936–3940).

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., & Wu, Y. (2021).

W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 244–250.

Cohen, A. S., Hong, S. L., & Guevara, A. (2010). Understanding emotional expression using prosodic analysis of natural speech: Refining the methodology. *Journal of Behavior Therapy and Experimental Psychiatry*, *41 2*, 150–157.

Cole, J., Hualde, I., Smith, C., Mahrt, T., & Napoleao de Souza, R. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *J. Phon*, *75*, 113–147.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*(1), 425–452.

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

Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory 2nd Edition. *Wiley Series in Telecommunications and Signal Processing*.

Davis, K., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, *24*, 637–642.

de Heer Kloots, M., & Zuidema, W. (2024). Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2.0. In *Proc. of Interspeech* (pp. 4593–4597). doi: 10.21437/Interspeech.2024-2490

Duarte, M., & Watanabe, R. (2018). *Notes on scientific computing for biomechanics and motor control.* `https://github.com/BMClab/BMC`. GitHub.

Elfeky, M. G., Moreno, P., & Soto, V. (2018). Multi-Dialectical Languages Effect on Speech Recognition: Too Much Choice Can Hurt. *Procedia Computer Science*, *128*, 1 - 8. (1st International Conference on Natural Language and Speech Processing) doi: https://doi.org/10.1016/j.procs.2018.03.001

El Zarka, D., Schuppler, B., Lozo, C., Eibler, W., & Wurzwallner, P. (2017). Acoustic correlates of stress and accent in Standard Austrian German. In S. Moosmüller, C. Schmid, & M. Sellner (Eds.), *Phonetik in und über Österreich, Veröffentlichungen zur Linguistk und Kommunikationsforschung: 31.* Vienna: ÖAW Austrian Academy of Sciences Press.

European Union. (2024). *Artificial Intelligence Act.* Retrieved 2024-10-23, from `https://www.artificial-intelligence-act.com` (Entered into force on August 1, 2024)

Facebook Research. (2022). *Fairseq Model (XLSR).* `https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec`. (Accessed: 2022-01-05)

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. Retrieved 2020-08-20, from `http://jmlr.org/papers/v20/18-760.html`

Forney, G. (1973). The Viterbi algorithm. *Proc. of IEEE*, *61*(3), 268–278. doi: 10.1109/PROC.1973.9030

Furui, S. (2009). Generalization problem in asr acoustic model training and adap-

tation. In *IEEE Workshop on Automatic Speech Recognition Understanding* (pp. 1–10). doi: 10.1109/ASRU.2009.5373493

Furui, S., Nakamura, M., Ichiba, T., & Iwano, K. (2005). Why is the recognition of spontaneous speech so hard? In V. Matoušek, P. Mautner, & T. Pavelka (Eds.), *Text, speech and dialogue* (pp. 9–22).

Gabler, P., Geiger, B. C., Schuppler, B., & Kern, R. (2023). Reconsidering read and spontaneous speech: Causal perspectives on the generation of training data for automatic speech recognition. *Information*, *14*(2). doi: 10.3390/info14020137

Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, *12*(2), 75 - 98. doi: https://doi.org/10.1006/csla.1998.0043

Gales, M., Knill, K., Ragni, A., & Rath, S. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*, 16–23. Retrieved from `https://eprints.whiterose.ac.uk/152840/`

Gales, M., & Young, S. (2007). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, *1*, 195–304. doi: 10.1561/2000000004

Geiger, B. C., & Schuppler, B. (2023). Exploring graph theory methods for the analysis of pronunciation variation in spontaneous speech. In *Proc. of Interspeech* (pp. 596–600). doi: 10.21437/Interspeech.2023-1398

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. of ICASSP* (Vol. 1, pp. 517–520). doi: 10.1109/ICASSP.1992.225858

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. of ICASSP* (Vol. 1, pp. 517–520).

Goldwater, S., Jurafsky, D., & Manning, C. D. (2008). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase ASR error rates. In *Proc. of ACL* (pp. 380–388). Columbus, Ohio: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P08-1044`

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA, USA: MIT Press. (`http://www.deeplearningbook.org`)

Google. (2023). *Reaper: Robust epoch and pitch estimator.* `https://github.com/google/REAPER`. GitHub.

Gopinath, R. A. (1998). Maximum likelihood modeling with Gaussian distributions for classification. In *Proc. of ICASSP* (Vol. 2, pp. 661–664).

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets. In *Proc. of ICML*.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. of ICML* (pp. 369–376). doi: 10.1145/1143844.1143891

Hadian, H., Sameti, H., Povey, D., & Khudanpur, S. (2018). End-to-end Speech Recognition Using Lattice-free MMI. In *Proc. of Interspeech* (pp. 12–16). doi:

10.21437/Interspeech.2018-1423

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proc. of the Sixth Workshop on Statistical Machine Translation* (pp. 187–197). Retrieved from `https://aclanthology.org/W11-2123`

Heckmann, M., Mikias, P., & Kolossa, D. (2014). The impact of word alignment accuracy on audio-visual word prominence detection. In *Speech Communication; 11. ITG Symposium* (pp. 1–4).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *ArXiv*, *abs/1606.08415*.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. doi: 10.1109/MSP.2012.2205597

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, *43*(1), 155–175. doi: https://doi.org/10.1016/j.specom.2004.01.006

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hornung, R., & Boulesteix, A.-L. (2022a). Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects. *Computational Statistics & Data Analysis*, *171*, 107460. doi: https://doi.org/10.1016/j.csda.2022.107460

Hornung, R., & Boulesteix, A.-L. (2022b). Supplementary material for "Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects". *Computational Statistics & Data Analysis*, *171*. (Supplementary Material 1) doi: https://doi.org/10.1016/j.csda.2022.107460

Hornung, R., & Wright, M. N. (2023). DiversityForest: Innovative complex split procedures in random forests through candidate split sampling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=diversityForest` (R package version 0.4.0)

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., ... Auli, M. (2021). Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. In *Proc. of Interspeech* (pp. 721–725). doi: 10.21437/Interspeech.2021-236

IPDS. (1997). *CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii.* Christian-Albrechts Universität zu Kiel.

Johnson, K. (2004). Massive reduction in conversational American English. In *Proc. in Spontaneous speech: Data and analysis. Proc. of the 1st session of the 10th international symposium* (pp. 29–54).

Kanda, N., Ye, G., Wu, Y., Gaur, Y., Wang, X., Meng, Z., ... Yoshioka, T. (2021). Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. In *Proc. of Interspeech* (pp. 3430–3434). doi: 10.21437/Interspeech.2021-102

Karanasou, P. (2013). *Phonemic variability and confusability in pronunciation modeling for automatic speech recognition* (Doctoral Thesis). Université Paris Sud - Paris XI.

Kessens, J. M., Strik, H., & Cucchiarini, C. (2002). Modeling pronunciation variation for ASR: Comparing criteria for rule selection. In *ITRW on Pronunciation*

Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA 2002) (pp. 18–23).

Khosravani, A., Garner, P. N., & Lazaridis, A. (2021). Modeling Dialectal Variation for Swiss German Automatic Speech Recognition. In *Proc. of Interspeech* (pp. 2896–2900). doi: 10.21437/Interspeech.2021-1735

Kim, J., & Kang, P. (2021). K-wav2vec 2.0: Automatic speech recognition based on joint decoding of graphemes and syllables. *ArXiv: abs/2110.05172*.

Kirch, W. (2008). Pearson's correlation coefficient. In *Encyclopedia of Public Health.* Dordrecht: Springer.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech and Language*, *45*(C), 326–347.

Klabbers, E., & Veldhuis, R. (2001). Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, *9*, 39–51.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. of Interspeech* (pp. 3586–3589). doi: 10.21437/Interspeech.2015-711

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *J. Acoust. Soc. Am*, *118*(2), 1038–1038.

Kohler, K. J., & Gartenberg, R. (1991). The perception of accents: F0 peak height versus F0 peak position. *AIPUK*, *25*, 219–294.

Kohler, K. J., Peters, B., & Scheffers, M. (2017). *The Kiel Corpus of spoken German - read and spontaneous speech. new edition, revised and enlarged.* Kiel, Germany: Kiel University. Retrieved from `http://www.isfas.uni-kiel.de/de/linguistik/forschung/kiel-corpus/`

Koiso, H., Amatani, H., Den, Y., Iseki, Y., Ishimoto, Y., Kashino, W., ... Watanabe, Y. (2022). Design and evaluation of the corpus of everyday Japanese conversation. In *Proc. of LREC* (pp. 5587–5594).

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., ... Usuda, Y. (2018). Construction of the corpus of everyday Japanese conversation: An interim report. In *Proc. of LREC.*

Kügler, F., Baumann, S., Andreeva, B., Braun, B., Grice, M., Neitsch, J., ... Wagner, P. (2019). Annotation of German intonation: DIMA compared with other annotation systems. In *Proc. of ICPhS* (pp. 1297–1301).

Laurent, A., Fraga-Silva, T., Lamel, L., & Gauvain, J. (2016). Investigating techniques for low resource conversational speech recognition. In *Proc. of ICASSP* (pp. 5975–5979).

Lehiste, I. (1970). *Suprasegmentals.* Cambridge: MIT Press.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, *10*, 707–710. Retrieved from `https://api.semanticscholar.org/CorpusID:60827152`

Linke, J., Garner, P. N., Kubin, G., & Schuppler, B. (2022). Conversational speech recognition needs data? experiments with Austrian German. In *Proc. of LREC* (pp. 4684–4691). Marseille, France: European Language Resources Association. Retrieved from `https://aclanthology.org/2022.lrec-1.500`

Linke, J., Geiger, B. C., Kubin, G., & Schuppler, B. (2024). What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures. *Computer Speech and Language*, 101738.

Retrieved from https://www.sciencedirect.com/science/article/pii/S0885230824001219

Linke, J., Kadar, M., Dosinszky, G., Mihajlik, P., Kubin, G., & Schuppler, B. (2023). What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers. In *Proc. of Interspeech* (pp. 5371–5375). doi: 10.21437/Interspeech.2023-951

Linke, J., Kelterer, A., Dabrowski, M. A., Zarka, D. E., & Schuppler, B. (2020). Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proc. Speech Prosody* (pp. 1000–1004). doi: 10.21437/SpeechProsody.2020-204

Linke, J., Kubin, G., & Schuppler, B. (2023). Using word-level features for prosodic prominence detection in conversational speech. In *Proc. of ICPhS.*

Linke, J., Steger, S., Steinwender, P., Kubin, G., Pernkopf, F., & Schuppler, B. (2025). Uncertainty prediction for prominence classification with chroma features. In *Proc. of icassp.*

Linke, J., Wepner, S., Kubin, G., & Schuppler, B. (2023). Using Kaldi for automatic speech recognition of conversational Austrian German. *ArXiv*, *abs/2301.06475*.

Lopez, A., Liesenfeld, A., & Dingemanse, M. (2022). Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing? In *Proc. of KONVENS* (pp. 135–143). Retrieved from https://aclanthology.org/2022.konvens-1.16

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *ArXiv*, *abs/1705.07874*.

Marsi, E., Reynaert, M., van den Bosch, A., Daelemans, W., & Hoste, V. (2003). Learning to predict pitch accents and prosodic boundaries in Dutch. In *Proc. of ACL* (pp. 489–496).

Mary, H. (n.d.). IARPA Babel Program.. Retrieved from https://www.iarpa.gov/research-programs/babel

Meghan, e. a., Glenn. (2013). *Gale phase 2 Arabic Broadcast Conversation Speech.* (Philadelphia: Linguistic Data Consortium (LDC))

Meyer, J. (2020). *Easy-kaldi.* https://github.com/JRMeyer/easy-kaldi. GitHub.

Mihajlik, P., Balog, A., Graczi, T. E., Kohari, A., Tarján, B., & Mady, K. (2022). BEA-base: A benchmark for ASR of spontaneous Hungarian. In *Proc. of LREC* (pp. 1970–1977). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.211

Mihajlik, P., Kádár, M. S., Dobsinszki, G., Meng, Y., Kedalai, M., Linke, J., ... Mády, K. (2023). What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced asr task? In *A Satellite Workshop of Interspeech (SIGUL 2023).*

Milde, B., & Köhn, A. (2018). Open source automatic speech recognition for German. In *Proc. of ITG* (pp. 251–255). Oldenburg, Germany.

Misra, H., Ikbal, S., Bourlard, H., & Hermansky, H. (2004). Spectral entropy based feature for robust ASR. In *Proc. of ICASSP* (pp. 193–196). IEEE.

Mixdorff, H., Cossio-Mercado, C., Hönemann, A., Gurlekian, J., Evin, D., & Torres, H. (2015). Acoustic correlates of perceived syllable prominence in german. In *Proc. of Interspeech* (pp. 51–55).

M. Mijwil, M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A.-H., & Bala, I. (2023). Chatgpt and the future of academic integrity in the artificial intelligence era: A new frontier. *Al-Salam Journal for Engineering and Technology*, *2*(2), 116–127. doi: 10.55145/ajest.2023.02.02.015

Mohamed, A., yi Lee, H., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1179–1210. Retrieved from `https://api.semanticscholar.org/CorpusID:248987289`

Moosmüller, S. (2007). *Vowels in Standard Austrian German. An Acoustic-Phonetic and Phonological Analysis.* University of Vienna: Habilitation Thesis.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., & Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech and dialogue* (pp. 424–431). Cham: Springer International Publishing.

Niebuhr, O. (2010). On the phonetics of intensifying emphasis in German. *Phonetica*, *67*, 1–29.

Niebuhr, O., & Winkler, J. (2017). The relative cueing power of f0 and duration in german prominence perception. In *Proc. of Interspeech* (pp. 611–615).

OpenAI. (2023). *Whisper Model (large-v2).* `https://github.com/openai/whisper`. (Release: openai-whisper==20230117)

ORF-TVthek: Broadcasts for the deaf and hard of hearing. (n.d.). `https://tvthek.orf.at/gehoerlosenservice`. (Accessed: 2022-07-20)

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 48–53). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-4009

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *Proc. of ICASSP* (pp. 5206–5210). doi: 10.1109/ICASSP.2015.7178964

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peterson, K., Tong, A., & Yu, Y. (2022). OpenASR21: The Second Open Challenge for Automatic Speech Recognition of Low-Resource Languages. In *Proc. of Interspeech* (pp. 4895–4899). doi: 10.21437/Interspeech.2022-10972

Popescu-Belis, A., Lalanne, D., & Bourlard, H. (2012). Finding information in multimedia meeting records. In *IEEE MultiMedia* (Vol. 19, pp. 48–57). doi: 10.1109/MMUL.2011.21

Povey, D. (2003). *Discriminaitve training for large vocabulary speech recognition* (Doctoral Thesis). University of Cambridge, Cambridge.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proc. of Interspeech* (pp. 3743–3747). doi: 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *Proc. of IEEE ASRU Workshop.* Retrieved from `http://infoscience.epfl.ch/record/192761`

Povey, D., et al. (2022). *Kaldi ASR TDNN Recipe Script.* `https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/local/chain2/tuning/run_tdnn_1i.sh`. (Accessed: 2022-01-03)

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. In *Proc. of Interspeech* (pp. 2751–2755). doi: 10.21437/Interspeech.2016-595

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. In *Proc. of Interspeech* (pp. 2757–2761). doi: 10.21437/Interspeech.2020-2826

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proc. of ICML* (Vol. 202, pp. 28492–28518). PMLR.

Rath, S. P., Povey, D., Veselý, K., & Černocký, J. (2013). Improved feature processing for deep neural networks. In *Proc. of Interspeech* (pp. 109–113). doi: 10.21437/Interspeech.2013-48

Reichel, U. D. (2012). PermA and Balloon: Tools for string alignment and text processing. In *Proc. of Interspeech.*

Reichel, U. D., & Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, 42–49.

Rogério Scalassara., P., Eugenia Dajer., M., Dias Maciel., C., & Carlos Pereira., J. (2008). Voice signals characterization through entropy measures. In *Proc. of the First International Conference on Bio-Inspired Systems and Signal Processing - Volume 2: BIOSIGNALS, (BIOSTEC 2008)* (pp. 163–170). SciTePress. doi: 10.5220/0001065401630170

Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 55–59). doi: 10.1109/ASRU.2013.6707705

Sato, Y., & Heffernan, K. (2020). Dialect clustering with character-based metrics: in search of the boundary of language and dialect. In *Proc. of LREC* (pp. 985–990). Marseille, France: European Language Resources Association. Retrieved from `https://aclanthology.org/2020.lrec-1.124`

Schmitt, B. J. B. (2018). *AMFM decompy documentation 1.0.8.* http://bjbschmitt.github.io/AMFM_decompy/.

Schuppler, B., Adda-Decker, M., & Morales-Cordovilla, J. A. (2014). Pronunciation variation in read and conversational Austrian German. In *Proc. of Interspeech.*

Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., & Pessentheiner, H. (2014). GRASS: The Graz corpus of Read And Spontaneous Speech. In *Proc. of LREC* (pp. 1465–1470).

Schuppler, B., Hagmüller, M., & Zahrer, A. (2017). A corpus of read and conversational Austrian German. *Speech Communication*, *94*, 62–74.

Schuppler, B., & Schrank, T. (2018). On the use of acoustic features for automatic homophone disambiguation in spontaneous German. *Computer Speech and Language*, *52*, 209–224.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proc. of Interspeech* (pp. 525–529). doi: 10.21437/Interspeech.2013-148

Schweitzer, A., Lewandowski, N., Duran, D., & Dogil, G. (2015). Attention, please! Expanding the GECO database. In *Proc. of ICPhS.*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423. Retrieved 2003-04-22, from `http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`

Shim, K., Choi, J., & Sung, W. (2022). Understanding the role of self attention for efficient speech recognition. In *International Conference on Learning Representations.* Retrieved from `https://openreview.net/forum?id=AvcfxqRy4Y`

Sriranjani, R., Karthick, B. M., & Umesh, S. (2015). Investigation of different acoustic modeling techniques for low resource Indian language data. In *Twenty First National Conference on Communications (NCC)* (pp. 1–5).

Stępkowska, A. (2012). Diglossia: A critical overview of the Swiss example. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, *129*, 199–209.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proc. of Interspeech* (pp. 901–904).

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, *29*(2), 225–246. doi: https://doi.org/10.1016/S0167-6393(99)00038-2

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. Retrieved 2013-06-28, from `http://www.biomedcentral.com.proxy.lib.uiowa.edu/1471-2105/9/307/abstract`

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., ... Carmiel, Y. (2020). WER we are and WER we think we are. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3290–3295). Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.295

Tamburini, F., & Caini, C. (2004). Automatic annotation of speech corpora for prosodic prominence. In *Proc. of LREC* (pp. 53–58).

Tamburini, F., & Wagner, P. (2007). On automatic prominence detection for German. In *Proc. of Interspeech* (pp. 1809–1812).

ten Bosch, L., Bentum, M., & Boves, L. (2023). Phonemic competition in end-to-end ASR models. In *Proc. of Interspeech* (pp. 586–590). doi: 10.21437/Interspeech.2023-1846

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *J. Acoust. Soc. Am*, *89*(4), 1768–1776.

Terken, J., & Hermes, D. J. (2000). The perception of prosodic prominence. In *Prosody: Theory and Experiment, studies presented to Gösta Bruce* (pp. 89–127). Dordrecht: Kluwer Academic Publishers.

Tian, J., Yu, J., Weng, C., Zou, Y., & Yu, D. (2023). Integrating lattice-free MMI into end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 25–38. doi: 10.1109/TASLP.2022.3198555

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, *52*(3), 201. doi: 10.1016/j.specom.2009.10.004

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *J. Acoust. Soc. Am*, *99*(6), 3782–3790.

Turnbull, R., Royer, A., Ito, K., & Speer, S. R. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Lang. Cogn. Neurosci.*, *32*(8), 1017–1033.

Tüske, Z., Saon, G., & Kingsbury, B. (2021). On the Limit of English Conversational Speech Recognition. In *Proc. of Interspeech* (pp. 2062–2066). doi: 10.21437/Interspeech.2021-211

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *ArXiv*, *abs/1807.03748*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Retrieved from `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

Veselý, K., Karafiát, M., & Grézl, F. (2011). Convolutive bottleneck network features for LVCSR. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 42–47). doi: 10.1109/ASRU.2011.6163903

Wagner, B., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., & D'Imperio et al., M. (2015). Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proc. of ICPhS*.

Wagner, P. (2005). Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proc. of Interspeech* (pp. 2381–2384).

Wahlster, W. (1993). Verbmobil: Translation of face-to-face dialogs. In *Proc. of Machine Translation Summit IV* (pp. 127–136). Kobe, Japan. Retrieved from `https://aclanthology.org/1993.mtsummit-1.11`

Walt, S. v. d., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22–30.

Wasserfall, S. (2020). *Automatic Speech Segmentation using Kaldi* (Unpublished master's thesis). Technical University Graz.

Wawra, D. (2014). *Job interview corpus. data transcription and major topics in corpus linguistics*. Berlin, Germany: Peter Lang Verlag. doi: 10.3726/978-3-653-04431-7

Wei, K., Guo, P., & Jiang, N. (2022). Improving Transformer-based Conversational ASR by Inter-Sentential Attention Mechanism. In *Proc. of Interspeech* (pp. 3804–3808). doi: 10.21437/Interspeech.2022-10066

Wepner, S., Schuppler, B., & Kubin, G. (2022). How prosody affects ASR performance in conversational Austrian German. In *Proc. of Speech Prosody* (pp. 195–199). doi: 10.21437/SpeechProsody.2022-40

Wright, R. (2006). Intra-speaker variation and units in human speech perception and ASR. In *ITRW on Speech Recognition and Intrinsic Variation (SRIV)* (pp. 39–42).

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The

Microsoft 2017 conversational speech recognition system. In *Proc. of ICASSP* (pp. 5934–5938).

Xu, J., Matta, K., Islam, S., & Nürnberger, A. (2021). German speech recognition system using DeepSpeech. In *Proc. of the 4th International Conference on Natural Language Processing and Information Retrieval* (p. 102–106). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3443279.3443313

Yamamoto, R. (2023). *Pyreaper: A python wrapper for REAPER.* `https://github.com/r9y9/pyreaper`. GitHub. (Release: version 0.0.8)

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying wav2vec2.0 to speech recognition in various low-resource languages. *ArXiv*, *abs/2012.12121*.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., ... Woodland, P. (2002). The HTK book. *Cambridge University Engineering Department*, *3*.

Zahorian, S., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking (Vol. 123) [Computer software manual].

Zeineldeen, M., Zeyer, A., Zhou, W., Ng, T., Schlüter, R., & Ney, H. (2020). A systematic comparison of grapheme-based vs. phoneme-based label units for encoder-decoder-attention models. *ArXiv*, *abs/2005.09336*.

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., ... Wu, Y. (2021). Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *ArXiv*, *abs/2109.13226*.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., ... Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *ArXiv*, *abs/2010.10504*.

Curriculum Vitae

	Academic Positions
09/2021 - 11/2021	Visiting Researcher at Idiap Research Institute, Martigny, Switzerland.
since 2019	University Assistant (Research and Teaching Associate) at Signal Processing and Speech Communication Laboratory (SPSC), Graz University of Technology.
	Education
since 2019	PhD student in Speech Signal Processing at Graz University of Technology.
2018 - 2019	Master's thesis on Sound Event Detection at Harman/Becker Automotive Systems.
10/2016 - 08/2019	Master's Degree in Electrical Engineering and Audio Engineering at Graz University of Technology and University of Music and Performing Arts Graz.
10/2012 - 09/2016	Bachelor's Degree in Electrical Engineering and Audio Engineering at Graz University of Technology and University of Music and Performing Arts Graz.

Research interests: Julian Linke is an interdisciplinary researcher since his Master's studies, combining the fields of acoustics, psychoacoustics, speech technology and speech science. His work aims at increasing our understanding of everyday communication. In more detail, he focuses on the development of automatic speech recognition systems and automatic tools for supra-segmental analyses, on the automatic detection of vocal fatigue, and on the integration of linguistic knowledge into speech technology. He further has substantial experience with data management, as well as experience with recording technology for the development of speech corpora.

Publications included in this thesis:

[A] Julian Linke, Saskia Wepner, Gernot Kubin, and Barbara Schuppler. (2023). Using Kaldi for automatic speech recognition of conversational Austrian German. *ArXiv* (abs/2301.06475).

[B] Julian Linke, Anneliese Kelterer, Markus A. Dabrowski, Dina El Zarka, and Barbara Schuppler. (2020). Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proc. of Speech Prosody* (pp. 1000–1004).

[C] Julian Linke, Gernot Kubin, and Barbara Schuppler. (2023). Using word-level features for prosodic prominence detection in conversational speech. In *Proc. of ICPhS* (pp. 3101–3105).

[D] Julian Linke, Bernhard C. Geiger, Gernot Kubin, and Barbara Schuppler. (2025). What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures. *Computer Speech and Language, Volume 90, 101738*.

[E] Julian Linke, Philip N. Garner, Gernot Kubin, and Barbara Schuppler. (2022). Conversational Speech Recognition Needs Data? Experiments with Austrian German. In *Proc. of LREC* (pp. 4684–4691).

[F] Julian Linke, Mate Kadar, Gergely Dosinszky, Peter Mihajlik, Gernot Kubin, and Barbara Schuppler. (2023). What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers. In *Proc. of Interspeech* (pp. 5371–5375).

Additional publications:

[AP1] Julian Linke, Sophie Steger, Philipp Steinwender, Gernot Kubin, Franz Pernkopf, and Barbara Schuppler. (2025). Uncertainty prediction for prominence classification with chroma features. In *Proc. of ICASSP* (in press).

[AP2] Karner Manuel, Julian Linke, Mark Kröll, Barbara Schuppler, and Bernhard C. Geiger. (2024). Towards Improving ASR Outputs of Spontaneous Speech with LLMs. In *Proc. of the 20th Conference on Natural Language Processing (KONVENS 2024)* (pp. 339–348).

[AP3] Peter Mihajlik, Yan Meng, Mate S. Kadar, Julian Linke, Barbara Schuppler, and Katalin Mády. (2024). On Disfluency and Non-lexical Sound Labeling for End-to-end Automatic Speech Recognition. In *Proc. of Interspeech* (pp. 1270–1274).

[AP4] Florian B Pokorny, Julian Linke, Nico Seddiki, Simon Lohrmann, Claus Gerstenberger, Katja Haspl, Marlies Feiner, Florian Eyben, Martin Hagmüller, Barbara Schuppler, Gernot Kubin, and Markus Gugatschka. (2024). VocDoc, what happened to my voice? Towards automatically capturing vocal fatigue in the wild. In *Biomedical Signal Processing and Control, 88* (pp. 105595–105606).

[AP5] Martin Hagmüller, Julian Linke, Simon Lohrmann, Florian Pokorny, and Barbara Schuppler. (2023). An acoustic analysis of vowels to predict voice changes in a long reading task. In *13th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications: MAVEBA* (pp. 23–26).

[AP6] Péter Mihajlik, Máté Soma Kádár, Gergely Dobsinszki, Yan Meng, Meng Kedalai, Julian Linke, Tibor Fegyó, and Katalin Mády. (2023). What kind of multi- or cross-lingual pre-training is the most effective for a spontaneous, less-resourced ASR Task?. In *A Satellite Workshop of Interspeech 2023 (SIGUL 2023)*.

[AP7] Anneliese Kelterer, Saskia Wepner, Julian Linke, and Barbara Schuppler. (2023). Points of maximum grammatical control–The prosody of a turn-holding practice. In *20th International Congress on Phonetic Sciences: ICPhS 2023* (pp. 3467–3471).

[AP8] Anneliese Kelterer, Sophie Christian, Saskia Wepner, Julian Linke, and Dina El Zarka. (2021). Prosodic cues to agreement and disagreement in "ja" and "nein" prefaces in Austrian German conversations. In *1st International Conference on Tone and Intonation: TAI 2021*.

[AP9] Julian Linke, Florian Wendt, Franz Zotter, and Matthias Frank. (2018). How the perception of moving sound beams is influenced by masking and reflector setup. In *VDT Tonmeistertagung*.

[AP10] Julian Linke, Florian Wendt, Franz Zotter, and Matthias Frank. (2018). How masking affects auditory objects of beamformed sounds. In *Fortschritte der Akustik, DAGA*.

[AP11] Franck Zagala, Julian Linke, Franz Zotter, and Matthias Frank. (2017). Amplitude panning between beamforming-controlled direct and reflected sound. In *Proc. of the AES 142nd Convention*.

Graz University of Technology