An analysis of prosodic boundaries across speaking styles in two varieties of German

Bogdan Ludusan[a,*], Barbara Schuppler[b,**]

[a]*Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany*
[b]*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria*

Abstract

Previous research has shown that differences in the marking of prosodic boundaries exist not only between varieties of the same language, but also between speaking styles of the same variety. Our aim is to gain insights into how these two factors interact in the case of German and Austrian German read and conversational speech. Analyzing four acoustic cues, representing durational and fundamental frequency (f0) measures, we observed that pause duration was the strongest cue to prosodic boundaries and that f0 reset was the weakest, in both varieties and across speaking styles. In read speech however, we noticed differences between the two varieties in the weighting given to the four cues. We then employed the four investigated acoustic cues for the annotation of the less-resourced variety, by examining ways of exploiting resources from the more well-resourced variety. Testing three automatic approaches for boundary detection, we obtained an overall high performance: 82.4% and 76.6% F-score for read and conversational speech respectively. Moreover, our results showed that training on more data from the well-resourced variety did not outperform the same system trained on less data from the target variety.

Keywords: prosodic phrase boundaries, conversational speech, read speech, automatic boundary detection, Austrian German, German

*Corresponding author
**Authors are listed in alphabetical order and contributed equally to this work
Email addresses: bogdan.ludusan@uni-bielefeld.de (Bogdan Ludusan), b.schuppler@tugraz.at (Barbara Schuppler)

1. Introduction

1.1. Prosody and speaking style

The study of prosodic variation and its functions has received much attention in speech research, with a methodological focus on controlled production experiments and read speech corpora (e.g., Volín et al. 2014; Luthern and Clopper 2015; Kim and Tilsen 2020). This methodological choice comes with the advantage that certain contexts of interest can be elicited and that various linguistic structures can be controlled for (e.g., El Zarka et al. 2019). With a general increase in the interest in phenomena occurring in conversational speech, investigations of prosodic variation in conversational speech have also become more prevalent in the speech community (e.g., Fuchs et al. 2010; Braun et al. 2020). The study of conversational speech, however, comes with the disadvantage that it requires more materials for the analysis and additionally more complex statistical and acoustic analysis methods in order to cope with the extensive variation present in this speaking style (e.g., Gubian et al. 2009; Ward 2019).

Yet, there seems to be no consensus in the literature when comparing prosodic variation across different speaking styles. While some studies find similarities between their controlled and more conversational materials (e.g., the analysis on the prosody of rhetorical and information-seeking questions in German by Braun et al. (2020)) and thus suggest that controlled experiments lead to generally valid conclusions, others obtain different, even contradictory, results. For instance, Sertling Miller (2007) investigated French prosody in Swiss and French speakers and found that whereas in read speech, Swiss and French speakers differ only slightly in terms of speaking rate and intonational patterns, in conversational speech these differences increase significantly.

The realization of prosodic phenomena across speaking styles has been examined also in Austrian German. Soukup (2007) performed an investigation of different styles (ranging from formal Standard Austrian German to colloquial/dialectal Austrian German) and reported that the styles do not differ in terms of intonation and prosody. These findings reflect those of Feizollahi and Soukoup (2011), showing that listeners were not able to distinguish dialect from standard in sentences where the segmental information

was filtered out. Nevertheless, significant differences in prosody were found between varieties of German (Hagmüller, 2001), which allowed the successful classification of these varieties based solely on prosodic features.

The outcomes of these studies highlight the importance of also taking into account the role of the speaking style when investigating prosodic phenomena. In this paper we focus on *speaking style* differences between read and conversational speech. The latter speaking style represents the type of speech used by adults in a normal conversation and may have different levels of spontaneity, depending on several factors, such as the context of the interaction or the familiarity of the speakers.

1.2. Prosodic boundaries

One of the main prosodic phenomena, prosodic phrasing, is involved, among others, in helping to parse the continuous speech into sentences (Cutler et al., 1997). The edges of prosodic phrases, known as prosodic boundaries, have been shown to be marked by several acoustic cues (e.g., pause duration, nucleus duration, pitch reset, etc.), both in read and conversational speech (e.g., De Pijper and Sanderman 1994; Peters 2003).

A salient cue for the presence of a prosodic boundary is the existence of a pause in the speech signal. Evidence to support this has been found in numerous linguistic studies that investigated the relationship between pause and boundary perception (e.g., Mo and Cole 2010), as well as in perception studies (e.g., Männel and Friederici 2016). The important role that pauses play in the marking of prosodic boundaries was also highlighted by the findings of Petrone et al. (2017), showing that adult German listeners give categorical responses for prosodic boundaries in the case of pauses, while more gradual transitions were observed with f0 and final lengthening cues. Compared to conversational speech, studies have shown that there are fewer, shorter, and more regular pauses in read speech (Silverman et al., 1992; Megyesi and Gustafson-Čapková, 2002; Wang et al., 2008; Sadat-Tehrani, 2017).

Final lengthening is another well studied phenomenon occurring in the vicinity of phrase boundaries, both in read and in conversational speech (e.g., Beckman and Edwards 1990; Fuchs et al. 2010; White et al. 2010; Holzgrefe-Lang et al. 2016). It affects the rhyme of syllables preceding a prosodic boundary, by means of which the corresponding speech segments are lengthened compared to when they are found in a phrase-medial position. Although the effect of speaking style on final lengthening has not been explicitly studied, the results of Church et al. (2005) seem to suggest that final lengthening

may be stronger in read than in conversational speech[1].

In addition to a longer segment duration before a phrase boundary, domain initial speech sounds may be produced with a stronger contact between articulators, with a decreased co-articulation to the previous context and a lengthening of the consonant duration in domain-initial position (Fougeron and Keating, 1997; Cho and Keating, 2009). Thus, the syllable onset following a prosodic boundary may be longer than the same syllable onset, produced phrase-internal. To our knowledge, no study so far has investigated differences in initial strengthening between different speaking styles.

Finally, f0 reset has been found to be an acoustic cue marking the presence of boundaries of major prosodic phrases, across various languages (e.g., Swerts 1997; Yang and Wang 2002; Vaissière 2005). Regarding differences between speaking styles, Swerts et al. (1996) reported that read speech exhibited a much stronger resetting as compared to conversational speech.

For Austrian German, only a limited number of studies have focused on prosody so far (e.g., Moosmüller and Brandstätter 2014; Moosmüller 2015; El Zarka et al. 2017; Leykum 2019; Siddins and Mennen 2019), and even fewer have investigated the acoustic marking of prosodic boundaries (Ulbrich, 2006; Schuppler and Ludusan, 2020). Ulbrich (2006) compared prosodic phrasing in read broadcast news corpora from the three German standard varieties (Germany, Austria, Switzerland). The analysis focused on the following measures: number of intra-sentential boundaries, number and duration of intra-sentential pauses, f0 reset, and phrase initial and phrase final syllable duration. It found that, whereas f0 reset and pausing was used equally among the different varieties, speakers from Germany followed the punctuation marks more closely (syntax-prosody relationship) than speakers from Austria and Switzerland and had the lowest change in speech rate across boundaries. The study by Ulbrich (2006), however, did not compare these acoustic cues in boundary-adjacent vs. phrase-medial position as we do here, and as we did in our previous study on read speech (Schuppler and Ludusan, 2020). In that study, we analysed four acoustics cues: pause existence, nucleus duration, nucleus-onset-to-nucleus-onset duration and f0 reset. It revealed that all four cues mark prosodic boundaries both in Austrian German and in German, with differences in the marking of boundaries between vari-

[1]However, the syllable structure was not controlled for in this study, which might have an impact on its findings.

eties for the nucleus-onset-to-nucleus-onset and the f0 reset features. Other investigations have looked at prosodic boundaries in Austrian German with the goal of modelling boundaries in speech synthesis systems (Neubarth et al., 2000; Apel et al., 2004; Pirker and Neubarth, 2003). However, since the collected and prosodically annotated data consisted of read speech of a single speaker, these studies could not offer a more general view of the phenomenon.

To our knowledge, for Austrian German, there has been no study comparing the acoustics of boundary-adjacent vs. phrase-medial syllables in spontaneous conversations. We aim to fill this gap, by analysing the acoustic marking of prosodic boundaries in Austrian German conversational speech, comparing it to similar materials from the language variety spoken in Germany and to read speech data.

1.3. Automatic detection of prosodic boundaries

As we have previously mentioned, conducting prosodic investigations on conversational materials requires a larger quantity of data than those employing read speech, due to the lack of control on the elicitation method and to the higher degree of variability present in the speech signal. One of the most time-consuming, and thus also most expensive tasks of corpus creation is their manual annotation of different annotation layers (i.e., orthographic, phonetic, prosodic). Once a critical amount of material for one language has been gathered, the development of automatic transcription tools is possible, and additional resources for that language can be created with relatively low effort. However, this is not the case for low-resourced languages, and for less-resourced varieties of well-resourced languages as, for instance, Austrian German. Although for some applications, a less-resourced variety may exploit existing resources from a well-resourced variety, this may not apply across the board. One such example is the automatic phonetic segmentation system MAUS (Kisler et al., 2017), which was trained on German data and works relatively well on read Austrian German, but less so for conversational Austrian German (Schuppler et al., 2014b).

Out of the possible annotation levels, one of the most time consuming is the prosodic layer, which also reaches lower inter-labeller agreements in conversational than in read speech. Previous approaches for automatic prosodic boundary detection use a variety of sources of information for positing boundaries, from acoustic, to lexical and even syntactic (e.g., Ananthakrishnan and Narayanan 2008; Christodoulides et al. 2017). Since newly created speech

databases normally have orthographic transcriptions (from which the phonetic segmentation may be derived by means of forced-alignment systems), we focus here on boundary detection approaches which make exclusive use of acoustic information.

For German, a number prosodic detection systems based on acoustic features have been proposed (Strom, 1995; Batliner et al., 2001; Braunschweiler, 2003; Soto et al., 2013; Stehwien et al., 2020). They normally include pause and nucleus duration, fundamental frequency (f0) measures, as well as speech intensity features, extracted either at the syllable or at the word level. Braunschweiler (2003) detected automatically GToBI labels (including boundary tones) on read and conversational speech, discovering 56% of the manually established labels. Employing similar materials, from conversational dialogues, Strom (1995) reports classification scores of up to 67%, while the approach presented in Batliner et al. (2001) reached a detection performance of 76%. Finally, employing information also regarding word boundaries and more powerful learning algorithms, Soto et al. (2013) and Stehwien et al. (2020) report classification F-scores of up to 92% for German broadcast news.

The only approach for automatically detecting prosodic boundaries in Austrian German is a study we conducted previously (Schuppler and Ludusan, 2020), in which we compared the performance of an acoustics-only detection system (Ludusan and Dupoux, 2014) on read speech uttered by speakers from Northern Germany and from Austria. The observed differences in how the acoustic cues were used in marking boundaries in the two varieties, also translated into different detection performances. A lower detection rate was obtained in Austrian German – 0.215 area under the receiver operating curve (AUC), compared to 0.308 AUC, in Northern German.

Given that differences may exist in how boundaries are marked in conversational speech, compared to read speech, and that these differences may affect the functioning of the automatic detection systems, we would like to extend our analysis also to Austrian German conversational speech data.

1.4. Current study

The current study will analyse how strongly different acoustic cues contribute to the marking of prosodic boundaries in read and conversational speech, in two varieties of German: German spoken in the Northern part of Germany (subsequently referred to as Northern German) and the variety spoken in Eastern Austria (subsequently referred to as Austrian German). We expect to obtain similar findings to those reported by Sertling Miller

(2007), who examined different varieties of French, also in our German data: to see larger differences between varieties in conversational speech than in read speech. Four acoustic cues will be investigated in a crossed design (2 varieties x 2 speaking styles), cues associated with a number of phenomena occurring in the vicinity of prosodic boundaries (Ludusan and Dupoux, 2014). They are: pause duration, the duration of the syllabic nucleus, the distance between the onset of the current syllable and that of the following syllable and the size of the f0 reset. In particular, we aim at identifying not only the significance of these cues in signalling prosodic boundaries, but also at determining their importance/weighting in the different speaking styles. Given that articulation rate has been shown to affect the realization of prosodic structure Trouvain and Grice (1999) and to differ significantly across speaking styles (e.g., Markó and Kohári 2015 for read vs. spontaneous Hungarian) and across varieties (Verhoeven et al. 2004 for Dutch vs. Flemish), we will take into account this factor in our analyses.

We hold that this analysis will provide a better understanding of prosodic boundaries in the two German varieties, by establishing whether the analysed cues are used for marking the boundary/non-boundary distinction. However, we would also like to ascertain the helpfulness of the four cues for the automatic discrimination of prosodic boundaries. It may be that not all of them are useful in a detection system as, for instance, a particular cue might not bring in new information compared to other cues or that the number of cases in which it would help the detection process might be lower than the number of cases it might hinder detection. Therefore, we will test also various approaches for automatic boundary detection, based on the features investigated in this study. With this analysis we also aim to facilitating the ongoing prosodic annotation process of GRASS, the first large scale database for read and conversational in Austrian German (Schuppler et al., 2014c, 2017). As only a small portion of the corpus has so far been manually annotated for prosodic boundaries, we hope to be able to use the automatic approaches as a first pass tool in a semi-automatic approach for annotating the rest of the corpus.

The current study extends our previous work on read speech (Schuppler and Ludusan, 2020), by including in the acoustic cue analysis also comparable conversational materials in the two varieties, by testing three detection systems on conversational data, as well as by examining whether training the systems on data from one variety generalizes well enough to the other variety.

2. Materials

2.1. Kiel corpus

The Kiel Corpus of Spoken German (Kiel corpus; Kohler et al. 2017) contains read and conversational speech produced by speakers mainly from Northern Germany. All the materials have been orthographically transcribed and phonetically annotated, with the majority of the recordings having also prosodic annotations. The prosodic annotations were created following the Kiel Intonation Model (KIM; Kohler 2006) and included annotations for prosodic boundaries, sentence accent and lexical stress.

The corpus contains materials uttered in two speaking styles: read and conversational speech. The read speech component (referred to as *Kiel_R* in the remainder of the paper) consists of around five hours of recordings, representing sentences and stories read by 53 speakers (26 females, 27 males). The conversational component includes two datasets. First, we used data from the "appointment-making-scenario" (referred to as *Kiel_A* in the remainder of the paper), which contains approximately five hours of speech from 43 speakers (22f, 21m) who were paired in dialogues with the task of making appointments. In this scenario, participants were only able to speak while holding a button pressed, which was also an action that blocked the channel of the other speaker, thus effectively avoiding any overlapping speech. The the speech produced in this scenario is thus more spontaneous, in terms of pronunciation and syntax than the speech of the read speech component, however, other aspects of natural conversational speech (e.g., related to turn taking) are not present in the data. Second, we also included in our analysis the data belonging to "video-task-scenario", which contains dyadic conversations on the topic of the "Lindenstraße" German television series (therefore abbreviated as *Kiel_L* in the remainder of the paper). Similar, but non-identical, video materials from the television series were presented separately to two subjects, after which they were asked to discuss the differences and similarities of the materials they had seen. In contrast to *Kiel_A*, in this scenario subjects knew each other and did not need to press a button in order to speak, leading to more natural and more spontaneous conversations than in the *Kiel_A* recordings. In both *Kiel_A* and *Kiel_L* speakers were sitting in separate rooms. Even though there are German corpora available with a higher degree of spontaneity in the conversations (e.g., Schweitzer et al.

2015), the Kiel corpus is the only one that comes with manually created prosodic annotations.

2.2. GRASS corpus

The Graz Corpus of Read and Spontaneous Speech (GRASS corpus; Schuppler et al. 2014c) contains read (6 hours) and conversational speech (19 hours) from a total of 38 speakers (19f, 19m) from the eastern provinces of Austria. Unlike the Kiel corpus, the speakers of the read and conversational speech component were identical. As language use in conversational Austrian German varies strongly with educational level, social background and region, speakers were chosen who were born in eastern Austria, had been living in either Graz or Vienna during their adulthood and either already had a university-level education or were at that time pursuing university-level education. The 19 speaker pairs were family members, friends, couples or colleagues, who had known each other for years. For the conversational component (referred to as *GRASS_C* in the remainder of the paper), they were recorded with head-mounted microphones, while sitting together in a recording studio for one hour, without interruption. There was no restriction in terms of chosen topic or speaking behavior, leading among other issues to a high degree of pronunciation variation (Schuppler et al., 2014a), frequently occurring overlapping speech and laughter (Schuppler et al., 2017). After participating in the conversational speech recording, speakers were asked to read short stories and selected isolated sentences, which made up the read component of the GRASS corpus (referred to as *GRASS_R* in the remainder of the paper). The materials recorded in the read part of the GRASS corpus were taken from the Kiel corpus.

Phonetic Annotation. The read speech component of GRASS was automatically segmented using MAUS (Kisler et al., 2017) and the subset used for this study was subsequently corrected manually by a phonetically trained annotator. As MAUS does not come with models suitable for Austrian German conversational speech, the conversational speech component of GRASS was automatically segmented using a KALDI-based forced alignment system currently under development at SPSC Laboratory (Wasserfall, 2020). It is based on 35 rules reflecting typical Austrian German pronunciation, including reductions and dialectal variants. All utterances used in this study were subsequently corrected manually by the authors. We thus believe that

by following this procedure, the segmental annotations of GRASS have a comparable quality to those found in the Kiel corpus.

Prosodic Annotation. At the prosodic level, a subset of GRASS was annotated manually for prosodic boundaries, following KIM, and thus following the same criteria as for the prosodic annotations of the Kiel corpus. Silence-intervals were defined as intervals without speech nor laughter, inbreaths and outbreaths were considered as part of silence. The guidelines suggested in Skarnitzl and Machac (2011) were followed for placing segment boundaries between words and phrases. Each utterance of the read speech component was annotated by one phonetically trained annotator and checked by two other annotators. The spontaneous conversations were annotated by one trained annotator, self-corrected in a second pass and, then, checked by one other annotator. This procedure was chosen to guarantee a high annotation quality. Based on a small validation set of 269 words from 47 utterances, which were randomly presented to the annotators during the transcription process, we calculated Cohen's kappa for the decision whether a boundary should be placed after a word or not. The obtained inter-annotator agreement for all prosodic annotations (including prominence annotations) was good to high (Cohen's kappa: $0.76 - 0.81$).

Communicative function annotation. A subset of the conversational speech component of GRASS has been annotated at the level of inter-pausal units (IPUs) for communicative function (Schuppler and Kelterer, 2021). In the annotation process, each IPU was given a label corresponding to the communicative function of that interval of speech, as perceived by an expert annotator. The functions included are: hold, incomplete-hold, change, question, hrt (hearer response token), self-interruption and trail-off. The conversations were, also on this level, annotated by one linguistically and phonetically trained annotator, self-corrected, and then checked by one other annotator.

2.3. Materials used in this study

In order to have materials as similar as possible between the two varieties of German, we post-processed the two corpora as follows: As the read component of the GRASS corpus contains only one sentence per file, we cut the files corresponding to the read stories (which contain several sentences), from the Kiel corpus, into the respective sentences, to match those of the GRASS corpus. Then, the recordings belonging to the *Kiel_A* conversations

were divided in speech turns, defined by the speakers themselves by pressing the button prior to speaking. The *Kiel_L* component came pre-segmented, each recording containing one or two speaker turns (from one or both speakers), excluding backchannels. Since no more detailed turn annotations exist for this dataset, we employed it as it is. The conversational speech files in the GRASS corpus contain the entire conversation and, thus, we made use of the IPU segmentation and the communicative functions annotation to determine speaker turns. A new turn started after each IPU that was not labelled as "hold" or "incomplete-hold". Subsequently, all the conversations in the GRASS corpus were cut into the obtained turns.

From the Kiel corpus, we employed all read speech recordings and all conversational speech recordings with channel separation (VerbMobil core g and Video-Task scenario l), which had both phonetic and prosodic annotations. For the GRASS corpus, the read speech materials included in this study correspond to the recordings having both manually corrected segmental annotations and prosodic annotations. The same was true for the conversational materials that were considered, the only difference was that they also had to be annotated for IPU communicative function, as this information was necessary for cutting the files into turns.

As the cues analysed in this study require syllable-level information and neither of the two corpora include this level of annotation, we derived syllables from the word- and phone-level segmentation. We first determined the position of all syllable nuclei for each word. We considered all vowels, as well as the syllabic sonorants /l/, /m/ and /n/ as syllable nuclei. We then placed syllable boundaries in correspondence to the sonority valley between each two consecutive nuclei (Clements, 1990).

Table 1 shows a summary of the materials used for the analyses presented in this paper. For both read and conversational speech, there are more materials available in the Kiel than in the GRASS corpus, as the entire Kiel corpus has already been manually annotated. Only a small subset of the GRASS corpus has been annotated for prosody or communicative functions, and one of the aims of this study is to investigate approaches that would facilitate the prosodic annotations of the remaining parts of the corpus. With respect to the speakers recorded in the two corpora, there is one main difference: Whereas in the Kiel corpus the speakers of the read speech and the conversational speech are not the same (only one speaker appears both in the read speech and in the *Kiel_A* component, and none of the speakers of *Kiel_L* appear in the other two components), in the GRASS corpus the same

Table 1: Summary of the materials used in this study. It details the number of word tokens, syllable tokens, phrase boundaries and speakers present in each corpus/component. The column *art. rate* displays the average articulation rate (i.e., the mean number of realized syllables per second), computed over all the syllables in each dataset.

	#wrd tokens	#syl tokens	#phrase bound	# spkrs	art. rate mean	art. rate SD
Read speech						
Kiel_R (255 min)	31,362	49,791	6,483	53	5.30	0.84
GRASS_R (79 min)	8,159	13,215	1,968	38	5.09	0.76
Conversational speech						
Kiel_A (223 min)	36,506	53,895	9,190	30	5.28	0.81
Kiel_L (75 min)	13,934	18,570	3,882	12	5.51	0.74
GRASS_C (43 min)	7,784	10,140	2,346	20	5.68	1.09
All data (675 min)	97,745	145,611	23,869	132	5.33	0.84

speakers were recorded in both speaking styles. This is also the reason why the total number of speakers in Table 1 is not the sum of the number of speakers found in the different components.

3. Methods

3.1. Acoustic features

Four acoustic cues were included in this study, cues which represent duration and pitch-related measures. They represent the acoustic correlates of a number of phenomena associated with prosodic structure: pausing, final lengthening, initial strengthening and pitch reset (e.g., Fougeron and Keating 1997; Vaissière 2005; Fletcher 2010). Moreover, they have been previously considered in relation to prosodic boundary investigations, being the most studied correlates of these phenomena. The four cues (and their names used throughout this study) are:

- duration of the following pause (*Pause*)
- duration of the syllable nucleus (*Nucleus*)
- the nucleus-onset-to-nucleus-onset duration (*Onset*)

- f0 reset (*f0_Reset*)

The pause feature characterizes the realization of pausing and numerous studies have looked at the duration of pauses in the vicinity of boundaries (e.g., Mo and Cole 2010; Simon and Christodoulides 2016; Petrone et al. 2017). The nucleus and onset features are both affected by final lengthening, the phenomenon by which the duration of the speech segments preceding a phrase boundary are lengthened compared to the same phrase-medial segments. The nucleus duration was investigated with respect to prosodic boundaries in previous studies, such as Holzgrefe-Lang et al. (2016); Ludusan et al. (2016); Yoon et al. (2007). The onset feature also captures changes due to initial strengthening, as segments following a boundary are pronounced with an increased articulatory effort, resulting in longer durations. Furthermore, it also includes the duration of the pause, if one follows the current prosodic boundary. Thus, the onset feature (Christophe et al., 2003; Ludusan and Dupoux, 2014) is a composite measure and captures the changes induced by the pausing, final lengthening and initial strengthening phenomena. Finally, since major phrase boundaries are usually associated with pitch resets we included the magnitude of this reset, similarly to, for instance Swerts (1997); Yang and Wang (2002); Kim (2019).

The values of these cues were then extracted for each syllable in our datasets, with the duration features being derived from the annotations supplied with the corpora. The f0 values were extracted using YIN (De Cheveigné and Kawahara, 2002), considering a pitch range between 60 and 500 Hz, and an interval between estimates of 1 ms (with the rest of the parameters using their default values). We have chosen YIN because it provides pitch estimation also for syllable nuclei having an irregular phonation, where other trackers may fail to return any pitch value.

Thus, for each syllable, the pause feature was represented by the duration of the pause following that syllable[2] (or having the value 0, if no pause followed the syllable). The pause was defined as an interval in which the speaker did not produce speech nor laughter. A pause might not be completely silent, as it may contain breathing sounds. The nucleus feature consisted of the duration of the syllabic nucleus, while the onset feature was the sum of the

[2]Previous works investigating the role of pauses in the marking of prosodic boundaries (e.g., Mo and Cole 2010; Petrone et al. 2017) used similarly low threshold values in their analyses.

duration of the syllabic nucleus, of the coda of the syllable, of the following pause, as well as the onset of the following syllable (if any of these components were missing, they were considered to have a value equal to 0). Lastly, f0 reset was computed as the difference between the mean f0 value of the following syllable nucleus and the mean f0 value of the current nucleus. If this value was lower than 0 (no pitch reset occurred), the feature was given the value 0.

All the features were then scaled between 0 and 1, on a sentence (for the read components) or a turn-basis (for the conversational components). The pause feature was obtained by truncating all pauses longer than one second, to one second. The truncation was performed in order to reduce the role of very long pauses that might appear within the conversational speech components. The f0 reset value was normalized by dividing each value of the same sentence/turn by the maximum f0 reset in the given sentence/turn. For the nucleus and the onset feature, as they might be prone to outliers (especially in the conversational components, because of very long syllabic nuclei due to hesitations or because of long turn-internal pauses), we transformed the values by dividing them by either the maximum value within the same unit (sentence/turn) or by the upper outlier limit, as computed using the Tukey method (see Equation 1), whichever the lowest. If, after the division by this value any of the resulting values are higher than 1, they are replaced by the value 1 (thus, the value corresponding to any of the instances of outliers, will not be higher than 1).

$$Outlier\_limit = Q3 + 1.5 * IQR \tag{1}$$

where IQR is the interquartile range ($IQR = Q3 - Q1$), Q1 is the first quartile and Q3 is the third quartile.

3.2. Statistical analyses

In order to investigate the contribution of the acoustic features to the marking of prosodic boundaries in Northern German vs. Austrian German and in read vs. conversational speech, we built linear mixed effects regression models with *Boundary* (Y, N) as the dependent variable and the previously mentioned acoustic measures *Pause* (duration), *Nucleus* (duration) and *f0_Reset* as independent variables. Since the nucleus-onset-to-nucleus-onset duration correlates heavily with *Pause* (for GRASS: r = 0.64, p<0.001; for Kiel corpus: r = 0.56, p<0.001) and *Nucleus* (for GRASS: r = 0.52, p<0.001;

for Kiel corpus: r = 0.56, p<0.001), the variable was orthogonalized and its residuals were added as the variable *Onset* to the models. We included also the continuous independent variable *Tempo* (i.e., articulation rate), which was calculated as the number of realized syllables (as given by the phonetic transcriptions) per second, determined on a per-sentence basis, for the read speech components, and on a per-turn basis, for the conversational speech components. Furthermore, the factors *Variety* (German, Austrian) and *Style*, which refer to the different components of the corpora used, as well as the random intercepts *Speaker* and *Syllable*, are taken into consideration.

For building the mixed effects logistic regression models, we used the `glmer()` function of the `lme4` package (Bates et al., 2015) in R (R Core Team, 2020) and performed the following steps: First, full models were build, including all independent variables and their interactions (two-way and three-way). Then, iteratively, non-significant predictors and interactions were removed as long as the model would still significantly improve given its AIC value, its degrees of freedom (Baayen, 2008; Levshina, 2015) and a model-comparison using the `anova()` function. Also random slopes as well as random intercepts were only added to the model when it improved given model-comparison using the `anova()` function. The significance threshold value was set at $\alpha = 0.05$ for all tests.

Next, we investigated the importance of the four acoustic cues in the two language varieties and the two speaking styles, by means of Random Forest classifiers used to predict the presence of a boundary based on these acoustic cues. This approach, of using Random Forest to rank the importance of features in marking linguistic structure, has already been used for prosodic phenomena (e.g., Ludusan et al. 2021, for prosodic prominence). We employed the implementation offered by the *randomForest* package (Liaw and Wiener, 2002) in R. A Random Forest classifier was run for 500 times, on each condition (component/speaking style combination). The overall contribution (or importance) of each feature was defined as the total decrease in node impurities (as measured by the Gini index) from splitting on that particular feature and was determined by averaging over all constructed trees (value returned by the function *importance* from the same R package). We then normalized, within each condition, by dividing the importance of each feature by the importance of the least important feature. Thus, we obtain for the least important feature a value of 1, while the values of the other features will reflect their importance with respect to the least important feature in that condition.

3.3. Prosodic boundary detection

As mentioned in the Introduction we are also interested in ways of facilitating the prosodic annotation of less-resourced varieties, by exploiting materials from well-resourced varieties. While the statistical analyses described in the previous section will help establish the role of each acoustic cue in the signalling of prosodic boundaries, it is likewise necessary to determine how much they actually help the automatic discrimination of boundaries. For this, we evaluate them in relation to three different detection systems.

The two corpora employed in the statistical analyses were also used for the automatic prosodic boundary detection experiments. For the GRASS corpus, the data corresponding to each of the two speaking styles was divided into into two parts: a training set composed of data from 4 speakers (2m, 2f) and a test set consisting of the remaining speakers (34 in the read part and 16 in the conversational part). The speakers were chosen as they appeared in both subsets and produced a number of boundaries close to the per-speaker subset average, in both styles. We considered a test set containing more data than the training set in order to ensure that our findings have higher generalization. This was made possible also by employing three detection systems which do not require much data for an optimum performance. A second reason for having a small train set was to compare the discrimination performance in a low-resource setting (as in the case of a less-resourced variety having a low amount of annotated data) with the case in which more training data is available, but from another variety. With regards to the Kiel corpus materials, its three sub-parts (Kiel_R, Kiel_A and Kiel_L) were used entirely as training sets.

The same four acoustic features were considered in the detection experiments: pause duration, nucleus duration, nucleus-onset-to-nucleus-onset duration and f0 reset. The prosodic boundary detection performance was evaluated using three classical measures: precision (the proportion of correctly classified boundary instances out of the total number of instances classified as being boundaries), recall (the proportion of correctly classified boundary instances out of the total number of boundaries in the dataset) and F-score, defined as being the harmonic mean between precision and recall.

We evaluated three algorithms here, encompassing three different learning/supervision approaches. The first one (RB), is a rule-based system, similar to the one proposed by Ludusan and Dupoux (2014). It computes, based on these four features, a syllable-based detector function. Since each feature

should have a high value for the syllable preceding a boundary, the algorithm looks for the local maxima (exceeding a certain threshold, k) of the obtained detector function and places prosodic boundaries after the corresponding syllable. This method has been previously employed for boundary detection (Ludusan and Dupoux, 2014; Schuppler and Ludusan, 2020) by summing up the values of the four features. We took a slightly different approach here, however, and instead of computing an overall function, we computed individual detector functions for each feature separately. The decision for a boundary was taken by combining the decisions of the individual functions by means of logical *or*. We then combined the feature as follows: Starting with the pause feature, a new feature was added to the system if the combined system performance (F-score) increased, compared with the system without that feature. Once the parameter k of each individual function and the best combination of features was determined on the training set, we used this configuration to run the algorithm on the test set.

The second algorithm (EM) uses an unsupervised learning paradigm. At training time, it tries to fit an equal number of Gaussian distributions to the expected number of categories, on the training data, by means of the Expectation Maximization algorithm. It uses no class information during training. At test time, the algorithm returns the probability of the given instances belonging to each of the clusters. The system was given the number of classes, two (boundary/non-boundary), and it was run for a maximum of 100 iterations, with a minimum allowable standard deviation of $1e-6$.

The last algorithm (NB) is a supervised classifier. It makes the same assumptions about the shape of the categories as the second algorithm and considers each input feature as being independent from one another. The parameters of the distributions are estimated by the classifier during training, which makes use of class labels. At test time, the posterior probability of each class is obtained by means of Bayes' formula. The algorithm returns the label corresponding to the class having the highest probability.

For the algorithms EM and NB, we used the implementation offered by the scikit-learn machine learning library (Pedregosa et al., 2011), namely mixture.GaussianMixture and naive_bayes.GaussianNB, respectively.

In order to determine the performance of the three systems on read and conversational Austrian German speech, we trained them/optimized their parameters on either the corresponding training set obtained from the GRASS corpus (matched condition) or on those from the Kiel corpus (mismatched conditions) and always tested on the GRASS test set. Having training data

from both varieties for each speaking style will allow us to evaluate whether successful transfer learning from a variety to another is possible and to compare performance. Thus, for each test set of the GRASS corpus (read, conversational) we trained on the corresponding speaking style either using data from the GRASS corpus or from the Kiel corpus. This was done for all algorithms, resulting in 15 cases (read x 2 training corpora x 3 algorithms + conversational x 3 training corpora x 3 algorithms).

4. Analysis of acoustic cues marking prosodic boundaries

We present here an analysis of how strongly the acoustic measures and their interactions contribute to marking prosodic boundaries (as given by the manual annotations of the data) in read vs. conversational speech in two varieties of German: Northern German and Austrian German. The mean values of the fours investigates acoustic cues across the two varieties and speech style conditions, for boundary and non-boundary positions, are presented in Table 2. Next, since earlier studies have shown an effect of articulation rate on pluricentric varieties and speaking styles, we take into account the articulation rate also in our analysis here (cf. Section 4.1). We then present our results from three mixed effects logistic regression models, (1) built on the GRASS corpus, to assess differences between read and conversational in Austrian German (cf. Section 4.2), (2) built on the Kiel corpus, to evaluate the role of speaking style in Northern German (cf. Section 4.3) and (3) on the conversational speech data from both corpora, to compare the two varieties in conversational speech (cf. Section 4.4). The question concerning how the two varieties differ in read speech has been addressed earlier in Schuppler and Ludusan (2020). We use mixed-effects logistic regression as it allows to directly relate it to the feature values and also to establish whether certain acoustic differences observed (e.g., pause duration is longer in boundary-adjacent than in phrase-medial position) are significant and have a large enough effect size. Furthermore, not only the main effects, but also the interactions between acoustic measures can be discovered. Since regression models do not provide feature ranking/importance information, we additionally use Random Forests to determine their overall role across speaking styles and varieties (cf. Section 4.5). At the end of this section, we will discuss the results from the different modelling techniques to draw our general conclusions (cf. Section 4.6).

Table 2: Mean values for the four acoustic features (normalized between 0 and 1), in phrase-medial (PM) and boundary-prior (BP) position.

	Pause		Nucleus		Onset		f0_Reset	
	PM	BP	PM	BP	PM	BP	PM	BP
GRASS corpus								
GRASS_R	0.000	0.512	0.472	0.722	0.467	0.888	0.175	0.385
GRASS_C	0.007	0.442	0.432	0.659	0.373	0.804	0.145	0.189
Kiel corpus								
Kiel_R	0.000	0.422	0.506	0.789	0.493	0.906	0.174	0.465
Kiel_A	0.001	0.233	0.437	0.720	0.427	0.805	0.121	0.207
Kiel_L	0.000	0.372	0.405	0.680	0.374	0.804	0.110	0.161

4.1. Articulation rates in read and conversational speech

Table 1 shows the mean and the standard deviation of the articulation rate, for each corpus and each speaking style. We built linear mixed models with *Tempo* as dependent variable, speaking style as predictor and *Speaker* as random effect, and found that in the GRASS corpus, the articulation rate in the conversational speech is significantly higher than in the read speech (β = 0.59, t = 48.49, p < 0.001). Furthermore, articulation rate showed to have higher variability (as shown by the standard deviation in Table 1) across analysis units (sentences/turns) and speakers in the conversational data than in the read speech. This result was to be expected given earlier studies showing that higher and more variable articulation rates occur in more spontaneous speaking styles (e.g., Markó and Kohári 2015 for Hungarian and Morrill et al. 2016 for native and non-native English).

In the Kiel corpus, the picture is more complex. The articulation rate is significantly lower in *Kiel_A* (β = -0.23, t = -33.17, p < 0.001) and in the read speech (β = -0.22, t = -31.16, p < 0.001) than *Kiel_L*, which is likely due to the more spontaneous nature of the materials contained in the *Kiel_L* component. However, the appointment-task-scenario was produced at a slightly slower (β = -0.01, t = -2.27, p < 0.05) and less variable articulation rate than the read sentences. These findings may be due to the format of the task employed in the Kiel corpus (making an appointment), which requires more planning than when casually chatting.

Finally, we observed that Austrians tend to read significantly slower than Germans (β =0.20, t = 32.62, p < 0.001) while producing a similar amount

of prosodic boundaries when reading: A boundary rate (defined as number of boundaries per second, excluding pauses) of 0.741 was found in the GRASS read speech, compared to 0.701 in the Kiel read speech component.

4.2. Results: GRASS corpus

This section compares how prosodic boundaries are realized in the two components of the GRASS corpus (i.e., read, conversational), allowing us to draw conclusions about differences in speaking styles in Austrian German. For our analysis, we built a mixed-effects logistic regression model for the GRASS corpus, with *Boundary* (phrase-medial (=intercept) vs.boundary-adjacent) as the dependent variable, and the acoustic features and speaking style as predictors. Table 3 shows the significant predictors (in the second column) separately for main effects and for the interaction terms with other predictors. The third column shows the *Estimate* of the predictor/interaction, which for the acoustic measures of interest (i.e., *Onset, Pause, Nucleus, f0_Reset*) equals the effect size, because all four measures range between 0 and 1). The fourth column contains the *z-value* for each predictor/interaction term of the model, which allows an equivalent to the *t-value* commonly seen in linear regression models. Finally, the fifth column shows the *p-value*, indicating whether the respective effect contributes significantly to the model.

The random effects of *Speaker* and *Syllable* significantly improved the model, reflecting the high variation among the 38 speakers and the impact of the syllabic structure (in terms of phone identities and neighborhoods) on the acoustic characteristics of prosodic boundaries. The acoustic cues *Nucleus, Onset* and *Pause* duration significantly mark prosodic boundaries in Austrian German, with *Pause* having the by far strongest effect size ($\beta = 29.69$), and *Onset* ($\beta = 6.64$), *Nucleus* ($\beta = 5.90$) having similarly high effect sizes for their main effects. The main effect of *f0_Reset* turned out to be not significant, its interaction with *Nucleus* duration, however, is highly significant: whereas for boundary-medial syllables, f0 reset increases with increasing nucleus duration, for boundary-adjacent syllables f0 reset is relatively constant independent of nucleus duration. Overall, based on the effect sizes of the main effects and of the interactions, f0 reset was shown to be the weakest and pause to be the by far strongest among the investigated acoustic cues to prosodic boundaries.

Three of the acoustic cues were shown to have highly significant interactions with *Style*: *f0_Reset*, *Nucleus* and *Pause*. The mean difference in

Table 3: The final model for prosodic boundaries in the GRASS corpus, as predicted by the four acoustic measures. For the factor *Style*, the value *GRASS_R* is in the intercept. The model includes the significantly contributing random intercepts *Speaker* and *Syllable*. $N = 23,355$, $AIC = 8611.2$.

GRASS corpus				
Type	Predictor	Estimate	z-value	p-value
	Intercept	-9.048	-19.17	< 0.001
Main effects:	Style(conv)	3.01	14.45	< 0.001
	Tempo	-0.29	-6.17	< 0.001
	Pause	29.69	11.14	< 0.001
	Nucleus	5.90	22.93	< 0.001
	f0_Reset	-0.11	-0.29	> 0.05
	Onset	6.64	6.68	< 0.001
Interactions:	Nucleus: f0_Reset	-1.54	-3.39	< 0.001
	Style(GRASS_C): Pause	-23.07	-8.92	< 0.001
	Style(GRASS_C): Nucleus	-2.23	-7.68	< 0.001
	Style(GRASS_C): f0_Reset	0.81	3.911	< 0.01
	Tempo: Pause	0.36	2.36	< 0.05
	Tempo: Onset	-0.43	-2.42	< 0.05

f0 reset between boundary-adjacent and phrase-medial syllables tends to be significantly smaller in conversational ($\Delta = 0.044$) than in read speech ($\Delta = 0.21$). In GRASS, prosodic boundaries are generally realized with significantly shorter nucleus duration in conversational than in read speech, and the difference between nucleus duration in phrase-medial vs. boundary-adjacent position is smaller in conversational speech (cf. Table 2). Also the mean difference in *Pause* duration between boundary-adjacent and phrase-medial syllables was significantly smaller in conversational ($\Delta = 0.43$) than in read speech($\Delta = 0.51$). We thus observe that overall, syllables in phrase-medial position are prosodically less distinguished from syllables in boundary-prior position in conversational than in read Austrian German.

Finally, the significant main effect of *Tempo* shows that, as expected, fewer boundaries are realized at higher utterance-level articulation rates. Its significant interaction terms indicate that with increasing articulation rates the effect of *Pause* duration in marking a boundary increases whereas the effect of *Onset* duration decreases.

4.3. Results: Kiel corpus

This section analyses how prosodic boundaries are realized in the three components of the Kiel corpus (i.e., Kiel_R, Kiel_A and Kiel_L), allowing us to draw conclusions about differences stemming from read vs. conversational speech in Northern German. Table 4 shows the final mixed-effects logistic regression model for all the data of the three components of the Kiel corpus. As for GRASS, the random effects of *Speaker* and *Syllable* significantly improved the model, and all acoustic features contributed significantly to the prediction of a boundary. With respect to the main effects, *Pause* had by far the highest effect size ($\beta = 39.59$), followed by *Onset* ($\beta = 8.98$) and *Nucleus* duration ($\beta = 5.45$), indicating that both onset and nucleus duration were significantly longer in boundary-prior than in phrase-medial position in all speaking styles of Northern German (cf. Table 2). Similar to GRASS, the main effect of *f0_Reset* turned out to be not significant.

There are significant interactions between the acoustic cues: The effect of *Onset* duration significantly decreases with increasing values of *f0_Reset* and with increasing *Pause* duration. Furthermore, this interaction with *Pause* duration is significantly greater for the less spontaneous speaking-styles Kiel_R and Kiel_A than for Kiel_L. When analyzing the function of pause duration conditioned by onset duration, we found that for boundary-adjacent syllables it this function has a smaller intercept and a steeper slope for Kiel_R than for Kiel_A and for Kiel_L.

Moreover, the model showed a significant main effect of *Tempo*, as earlier seen in the GRASS model, indicating that fewer boundaries are realized at higher articulation rates. *Tempo* also has significant interactions with *Onset* and *Pause* duration, with similar tendencies as observed in the GRASS corpus. In addition, *Tempo* has a significant interaction with *f0_Reset*, indicating that at higher articulation rates the effect of *f0_Reset* on marking a prosodic boundary increases.

Relevant for our analysis across speaking styles is that the model for the Kiel corpus contained highly significant interactions of all acoustic features with *Style*: Nucleus duration is significantly longer in read than in Kiel_L, but are similarly long in the two conversational components Kiel_L and Kiel_A. Onset duration also tends to be significantly greater for boundary-adjacent syllables in read speech (0.906) than in the two conversational components (Kiel_A: 0.805, Kiel_L: 0.805, cf. Table 2). The difference in onset duration between syllables in boundary-adjacent and phrase medial position, however, tends to be significantly larger in Kiel_L ($\Delta = 0.43$) than in the less

Table 4: The final model for prosodic boundaries in the Kiel corpus, as predicted by the four acoustic measures. For the factor *Style*, the value *Kiel_L* is in the intercept. The model includes the significantly contributing random intercepts *Speaker* and *Syllable*. $N = 122,256$, $AIC = 36557.1$.

KIEL corpus				
Type	Predictor	Estimate	z-value	p-value
	Intercept	-4.52	-27.50	< 0.001
Main effects	Style(Kiel_R)	-2.22	-15.84	< 0.001
	Style(Kiel_A)	-0.38	-3.48	< 0.001
	Tempo	-0.35	-15.15	< 0.001
	Pause	39.59	7.24	< 0.001
	Nucleus	5.45	36.08	< 0.001
	f0_Reset	-0.51	-1.28	> 0.05
	Onset	8.98	14.01	< 0.001
Interactions:	Style(read): Nucleus	0.80	0.81	< 0.001
style	Style(Kiel_A): Nucleus	-0.29	-1.67	> 0.05
	Style(Kiel_R): f0_Reset	-0.90	-4.57	< 0.001
	Style(Kiel_A): f0_Reset	0.98	5.87	< 0.001
	Style(Kiel_R): Pause	-26.62	-6.14	< 0.001
	Style(Kiel_A): Pause	-22.58	-5.31	< 0.001
	Style(Kiel_R): Onset	-0.80	-3.01	< 0.01
	Style(Kiel_A): Onset	-0.78	-3.53	< 0.001
Interactions:	Tempo: Pause	2.23	3.21	< 0.01
articulation rate	Tempo: Onset	-0.84	-7.52	< 0.001
	Tempo: f0_Reset	0.16	2.35	< 0.05
Interactions:	f0_Reset: Onset	-0.66	-3.35	< 0.001
acoustic cues	Pause: Onset	-45.60	-3.75	< 0.001
Three-way	Style(Kiel_R): Pause: Onset	96.26	7.15	< 0.001
interactions	Style(Kiel_A): Pause: Onset	22.73	1.79	> 0.05
	f0_Reset: Onset: Tempo	0.68	2.15	< 0.05

spontaneous speaking styles Kiel_A ($\Delta = 0.38$) and Kiel_R ($\Delta = 0.41$).

Similar conclusions regarding speaking style can be drawn also with respect to f0 reset, as seen from the mean values reported in Table 2. On average, with increasing spontaneity, f0 reset in boundary-prior position decreases

(Kiel_R: 0.465 > Kiel_A: 0.207 > Kiel_L: 0.161) and the difference in f0 reset between boundary-prior and phrase-medial position decreases (Kiel_R: Δ= 0.291 > Kiel_A: Δ= 0.086 > Kiel_L: Δ= 0.51). However, these results seem to contradict those illustrated in Table 4, for the interaction between *Style* and *f0_Reset*, in which different directions can be observed for read speech and Kiel_A, compared to Kiel_L (the intercept). We assume that this discrepancy might be due to the numerous significant two-way and three-way interactions of *f0_Reset* in the model. In order to assess this, we built another model with *Style* and *f0_Reset* as the only predictors, and *Speaker* and *Syllable* as random factors. The new model confirmed the trends seen from the average f0 reset values, revealing significant differences for *f0_Reset* across speaking styles. Considering Kiel_R as the intercept, the model showed the following significant interactions: *Style(Kiel_L)*: *f0_Reset*: $\beta = -1.21$, $z = -11.92$, $p < 0.001$; *Style(Kiel_A)*: *f0_Reset*: $\beta = -0.58$, $z = -8.76$, $p < 0.001$).

4.4. Results: conversational speech

We focus in this section on comparing how prosodic boundaries are realized in Austrian vs. Northern German conversational speech. In general, independent of the position of the syllable within a phrase (boundary-position or phrase-medial), Austrian speakers produced higher values of f0 reset than Northern German speakers (0.155 for GRASS_C vs. 0.133 for Kiel_A and 0.120 for Kiel_L) and approximately similarly long nucleus (0.484 for GRASS_C vs. 0.484 for Kiel_A and 0.460 for Kiel_L) and onset durations (0.472 for GRASS_C vs. 0.490 Kiel_A and 0.461 for Kiel_L). Austrian speakers produced also significantly longer pauses, both on average over all tokens (0.106 in GRASS vs. 0.0395 in Kiel_A and 0.0751 in Kiel_L) and also in boundary-adjacent position (0.442 in GRASS_C vs. 0.233 in Kiel_A and 0.372 in Kiel_L). Given the longer pause duration one might conclude that Austrians speak slower than Northern Germans in conversational speech. In respect to articulation rate, however, we observed that independent of whether syllables were boundary-adjacent or phrase-medial, in conversational speech, Austrian speakers produced higher articulation rates than the Northern German speakers (5.68 for the GRASS_C vs. 5.28 for Kiel_A and 5.51 for Kiel_L, cf. Table 1). This result, however, might not be related to the regional background of the speakers, but rather to the degree of spontaneity caused by the three different conversational setups, showing higher articulation rates the more spontaneous the conversations are (free conversations in GRASS > video-task in the Kiel corpus > appointment-task in Kiel corpus).

Table 5: Model for prosodic boundaries in the conversational speech components of the Kiel and GRASS corpus as predicted by the four acoustic measures. For the factor *Corpus*, the value *GRASS_C* is in the intercept. The model includes the significantly contributing random variable *Syllable*. $N = 82,605$, $AIC = 32525.3$.

Conversational speech				
Type	Predictor	Estimate	z-value	p-value
	Intercept	-3.94	-15.37	< 0.001
Main effects	Corpus(Kiel_A)	-0.88	-3.04	< 0.01
	Corpus(Kiel_L)	1.30	3.56	< 0.001
	Tempo	-0.17	-4.12	< 0.001
	Pause	8.70	20.84	< 0.001
	Nucleus	3.96	23.91	< 0.001
	f0_Reset	-0.31	-1.98	< 0.05
	Onset	3.60	33.54	< 0.001
Interactions:	Corpus(Kiel_A): Tempo	-0.11	-2.25	< 0.05
variety	Corpus(Kiel_L): Tempo	-0.42	-6.66	< 0.001
	Corpus(Kiel_A): Pause	9.44	9.98	< 0.001
	Corpus(Kiel_L): Pause	23.09	8.92	< 0.001
	Corpus(Kiel_A): Nucleus	1.19	6.36	< 0.001
	Corpus(Kiel_L): Nucleus	1.31	6.11	< 0.001
	Corpus(Kiel_A): f0_Reset	1.53	8.74	< 0.001
	Corpus(Kiel_L): f0_Reset	0.40	1.83	> 0.05
Interactions:	Pause: Onset	11.93	8.63	<0.001
acoustic cues	Pause: f0_Reset	-0.08	-0.08	> 0.05
	f0_Reset:Onset	-2.14	-7.92	< 0.001
Three-way	Corpus(Kiel_A): Pause: f0_Reset	5.38	2.05	< 0.05
interactions	Corpus(Kiel_A): Pause: f0_Reset	-2.80	-0.37	> 0.05

For read speech, we observed a different tendency, i.e., that Northern Germans have a higher articulation rate than Austrians (5.30 for Kiel_R vs. 5.09 for Grass_R) and also produce shorter pauses at boundary-adjacent position than Austrians (0.512 for GRASS_R vs. 0.422 for Kiel_R).

In order to investigate how the acoustic cues mark prosodic boundaries in conversational Northern German vs. Austrian German, we built a mixed effects logistic regression model on the conversational speech data from the

Kiel and GRASS corpus, and its findings are shown in Table 5. In line with the models presented in the previous sections, all four acoustic cues contributed significantly to the marking of a boundary. Whereas in the earlier presented models including read speech, *Pause* duration was shown to have by far the highest effect size on the main effect compared to the other acoustic cues, on conversational speech its effect is not that much larger than the one by *Nucleus* and *Onset* duration. But differently from previous models, this one shows a significant main effect of *f0_Reset*.

Next, we focus on the significant interactions between the four acoustic cues and the *Corpus* predictor. The acoustic measures *f0_Reset*, *Pause* duration and *Nucleus* duration proved to be variety dependent, all three measures having a stronger effect in predicting a prosodic boundary in the Northern German data than in the Austrian one. However, *Onset* duration exhibits no such variety dependent behaviour (a result also mirrored by the mean values given in Table 2). Furthermore, there were also significant three-way interactions with *Corpus*: The difference in mean values for f0 reset of boundary-adjacent and phrase-medial syllables were significantly higher for conversations produced by the Northern German than for the Austrian speakers (Kiel_A: $\Delta = 0.105$ and Kiel_L: $\Delta = 0.052$ vs. GRASS_C: $\Delta = 0.044$). The higher difference observed for in Kiel_A (see, for comparison, the values obtained for read speech, Kiel_R: $\Delta = 0.291$ vs. GRASS_R: $\Delta = 0.209$) is yet another indicator, besides the already mentioned characteristics such as articulation rate and pause duration, of its less spontaneous style compared to the other two conversational datasets used in this study. In all speaking styles, read and conversational, we observed that Northern German speakers produce a larger f0 reset difference between phrase-medial and boundary adjacent syllables than Austrian speakers.

4.5. Overall importance of the acoustic features

Figure 1 shows the ranking obtained for each cue: pause (black), nucleus (dark gray), onset (light gray) and f0 reset (white), by means of the Random Forest analysis. We see that for each condition, the two most important features are the pause, followed by the onset, while f0 reset helped less in decreasing the node impurities than any other feature. All within-condition differences between each two features were found to be significant, when using Wilcoxon signed rank tests.

In both read and conversational speech, the features show the same ranking, but the actual importance values are more similar in conversational

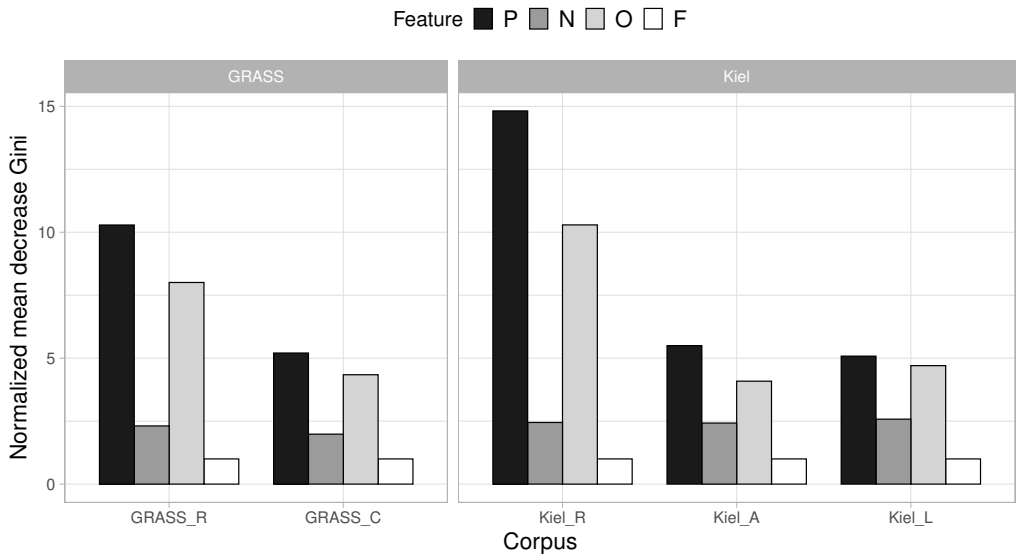

Figure 1: The overall importance of the four acoustic cues (pause (P), nucleus (N), onset (O), f0 reset (F)), in each of the two corpora and the two speaking styles, as given by a Random Forest classifier. The vertical axis represents the mean decrease in node impurities, normalized per corpus and style, by dividing each value by that of feature having the lowest decrease.

speech than in read speech, between the two language varieties. We notice a decrease in the importance of the pause and of the onset feature between the read and conversational conditions. This is probably due to the lower percentage of prosodic boundaries marked by a pause in the latter, compared to the former. Nevertheless, the decrease in importance for onset is lower than that observed for pause, suggesting that the other phenomena captured by this cue (final lengthening, initial strengthening) play a bigger role in conversational than in read speech.

4.6. Discussion

The aim of the analyses presented in this section was to investigate how strongly the acoustic measures pause, nucleus and onset duration as well as f0 reset contribute to marking prosodic boundaries in German vs. Austrian German read and conversational speech (i.e., four conditions). For this purpose, we first built mixed effects logistic regression models to investigate interactions between the acoustic cues and how they differ in the four con-

ditions. Subsequently, we performed a Random Forest analysis to estimate the overall importance of the four acoustic measures.

The regression models showed that prosodic boundaries are realized with longer onset and nucleus duration and that they are frequently marked by a pause in all conditions. Since we normalized all acoustic features between 0 and 1, we could compare how large differences were produced between bounday-adjacent and phrase medial syllables across the four acoustic cues. Among them, f0 reset exhibited the smallest difference between boundary-adjacent and phrase medial syllables. This is congruent with the outcomes of the Random Forest based analysis, where in all four conditions, the existence of a pause resulted to be the strongest predictor for prosodic boundaries, followed by onset duration, and with f0 reset being the weakest cue. We found that the pause also has complex interactions with the other acoustic cues, and that these interactions are variety-dependent. Similar results with respect to pauses were reported also in Schleef (2003). Investigating pauses in narratives in two other German varieties (Low German and Alemanic), the author observed that pause duration following higher level linguistic units tends to vary with the language variety.

Another aspect of our analysis focused on differences between the two considered speaking styles, read and conversational speech. We found that in conversational Austrian and Northern German, prosodic boundaries tend to be marked with lower onset and nucleus duration than in read speech, and that the differences in onset and nucleus duration between boundary-adjacent and phrase-medial syllables are significantly lower in conversational speech. In both varieties, boundaries in conversational speech are less frequently marked by a pause and also exhibit a lower f0 reset. The latter results are in line with the findings of Swerts et al. (1996), reporting a much stronger resetting in read, compared to conversational speech, in Swedish.

We chose to combine the regression analysis, which allows us to draw concrete conclusions about the variation and size of the acoustic measures, with a Random Forest analysis, which makes it possible to easily compare the relative importance of the investigated acoustic cues across the four conditions. It revealed that in Austrian German, the ranking of the acoustic cues is the same in read as in conversational speech, with the main difference being that pause and onset have a higher importance in read than in conversational speech. A similar importance ranking was also found in Northern German, with the only difference being that the weighting of the pause and onset features between read and conversational speech is larger than in the Austrian

data, suggesting a bigger role of these features in the former language variety.

Overall, our cue importance findings corroborate those of Petrone et al. (2017), in which adult German listeners gave more categorical responses for prosodic boundaries in the case of pauses, while more gradual transitions were observed with f0 and final lengthening cues. The fact that f0 reset was the weakest cue may be explained by the fact that the pitch reset phenomenon is mainly associated with the boundaries of higher level prosodic phrases (equivalent to intonation phrases). The edges of lower level phrases are often marked by an opposite f0 pattern and averaging across boundaries found at both levels would reduce (partly or entirely) the effect of f0 (see Ludusan et al. 2016 for a similar account with respect to Japanese prosodic boundary marking).

Conversational speech seems to exhibits a more similar acoustic features ranking than read speech. Thus, whereas the actual values of the acoustic measures in Austrian and German conversational speech are significantly different (as resulting from our regression analysis), the feature ranking/importance is the same. If anything, the two varieties seem to be more similar to each other in conversational rather than in read speech. This is in contrast to the findings of Sertling Miller (2007)[3], that showed larger differences in conversational than in read speech, between two varieties of French (spoken in France vs. spoken in Switzerland).

We also studied the variation of articulation rate in all conditions (at the utterance level in the read data and the turn level in the conversational data) and its relationship to the existence of prosodic boundaries. In the Kiel and the GRASS corpus there is the overall tendency of fewer boundaries and shorter onset and nucleus durations being realized at higher articulation rates. Furthermore, we observed that Austrians tend to read significantly slower than Germans while producing a similar quantity of prosodic boundaries when reading, resulting in a higher boundary rate. We found that whereas in the GRASS corpus, articulation rate tends to be higher and show more variability among speakers in conversational than in read speech, in the German data, the articulation rate in the less spontaneous component of the conversational speech (i.e., Kiel_A) is similar to that found in read speech, but with the more spontaneous component (i.e., Kiel_L) exhibiting

[3]Please note, however, that the study has investigated only the intonation patterns found at prosodic boundaries, without looking at the actual values of the acoustic cues.

a higher articulation rate, that is in the range of what has been found in the conversational speech of GRASS. We assume that this is due to the task-oriented setting of the Kiel_A component, in contrast to the casual topic-open conversations of the GRASS corpus.

Compared to previous investigations of pluricentric languages, our results for the articulation rate align well with those obtained on other languages, showing differences in speaking rates between varieties of the same language, in various speaking styles (Verhoeven et al. 2004, for conversational Dutch; Biadsy and Hirschberg 2009, for conversational Arabic; Yan and Vaseghi 2010, for read English; Velázquez 2010, for conversational Spanish;). It appears, however, that differences exist also between pluricentric languages, since no effect of speaking style was found across varieties of French (Schwab and Avanzi, 2015).

Our findings extend those of previous studies examining the marking of prosodic boundaries in varieties of German (Ulbrich, 2006; Schuppler and Ludusan, 2020). Compared to Ulbrich (2006), we did not limit our analysis only to the boundary position, but we considered also phrase-interval positions and we compared the two cases. Thus, we were able to determine which features are more discriminable for prosodic boundaries. With respect to the study presented in Schuppler and Ludusan (2020), we expanded our investigation to include also conversational materials with various degrees of spontaneity. We observed that, with increasing spontaneity of the speech style, the difference in the acoustic marking between boundary-adjacent and phrase-medial syllables decreased. Moreover, from a methodological point of view, our study shows that employing a second statistical methods not only confirmed the findings of the first analysis, but also yielded additional insights. Given the observed differences between the speaking styles in the two studied varieties, we can conclude that especially when aiming at investigating the differences between varieties of a pluricentric language language, the investigation of conversational speech may yield different results compared to those obtained from read data.

5. Prosodic boundary detection performance

5.1. Results

Table 6 presents the results of the boundary detection experiments. As mentioned in Section 3.3, the best combination of features for the RB algorithm was determined on the corresponding training set. For the read data,

Table 6: Automatic boundary detection performance on the read and conversational speech test sets of the GRASS corpus, while training the systems on data from the GRASS or Kiel corpus, respectively. Three algorithms were used (RB, EM and NB) and the classification precision, recall and F-score are reported.

	train	algorithm	precision	recall	F-score
Read speech	GRASS	RB	0.970	0.705	0.816
		EM	0.960	0.720	0.823
		NB	0.936	0.730	0.820
	Kiel	RB	0.931	0.738	0.824
		EM	0.960	0.720	0.823
		NB	0.954	0.724	0.823
Conv. speech	GRASS_C	RB	0.768	0.702	0.733
		EM	0.835	0.705	0.764
		NB	0.878	0.679	0.766
	Kiel_A	RB	0.793	0.696	0.742
		EM	0.661	0.723	0.691
		NB	0.843	0.700	0.765
	Kiel_L	RB	0.700	0.718	0.709
		EM	0.835	0.705	0.764
		NB	0.813	0.722	0.765

both in the matched and in the mismatched conditions, the pause by itself gave the best performance. In the case of conversational speech instead, the best combination included pause duration and onset duration, in all three investigated cases (one matched - GRASS_C, two mismatched - Kiel_A and Kiel_L).

In order to better understand our results we fitted a linear model with the per-speaker F-score results as the dependent variable and considering speaking style (read/conversational), experiment condition (matched/mismatched), algorithm (RB/EM/NB) as predictors. A type III ANOVA analysis of the fitted model revealed a significant main effects for speaking style ($p < 0.001$), with the performance on read speech being higher than on conversational speech. A marginally significant effect was obtained for the predictor algorithm, with the NB having a significantly higher performance than RB ($p < 0.05$).

Next, we compared the effect of the degree of spontaneity in the con-

versational speaking style. For this purpose, we fitted a linear regression model with the obtained F-scores, considering the dataset (GRASS_C, Kiel_A and Kiel_L) and the algorithm (RB/EM/NB) as predictors. The subsequent ANOVA analysis showed a significant effect only for the algorithm ($p < 0.05$) and the interaction between the dataset and algorithm ($p < 0.05$).

Post-hoc Wilcoxon tests were then performed to test all pairwise differences. The statistical significance of the between-speaking style differences were compared by means of per-speaker Wilcoxon rank sum tests. All the differences were found significant in the matched condition (train and test on the GRASS corpus), with the performance on read speech being higher than that for conversational speech (RB $p < 0.01$, EM and NB $p < 0.05$). In the mismatched condition (train on the Kiel corpus and test on the GRASS corpus), the tests revealed similar findings: RB and EM ($p < 0.001$) and NB ($p < 0.01$).

We then compared the differences between the matched and the mismatched condition, for each of the two speaking styles considered here, using per-speaker paired Wilcoxon tests. None of the differences were found to be significant, in read speech. For conversational speech, we analysed separately the difference between Austrian and German, considering we had two different datasets for the latter variety. When comparing the performance obtained on the GRASS corpus with that on the Kiel_A dataset, significant differences were found for RB ($p < 0.05$) and EM ($p < 0.001$). Considering the Kiel_L dataset, instead, only the RB algorithm showed a statistically significant difference ($p < 0.001$).

Finally, the differences between algorithms were checked by means of per-speaker Wilcoxon signed rank tests. The F-score in the matched read condition differed significantly between RB-EM ($p < 0.01$), with similar results also being obtained for the matched conversational condition (RB-EM $p < 0.001$ and RB-NB $p < 0.01$). In the mismatched read condition, the performance differences between EM-NB ($p < 0.01$) and RB-NB ($p < 0.05$) were found to be significant, while in the mismatched conversational condition the differences between RB-NB ($p < 0.001$) and EM-NB ($p < 0.001$) were significant.

5.2. Discussion

Our results indicate clear differences in automatic boundary detection between read and conversational materials, with all conditions exhibiting a higher performance for the former speaking style. This was to be expected,

given the more carefully pronounced speech in the read materials and the higher prevalence of pauses following a prosodic break in materials of this type (a similar observation having been made by Soto et al. (2013) for German read news). The important role of pauses marking boundaries in read speech can be also observed from the fact that in the matched condition, the pause feature by itself gave the best performance, thus confirming the finding of our previous analysis on the ranking of the acoustic features. Moreover, we also examined the effect of conversational data spontatenity on the detection performance, showing that no difference exists for automatic boundary detection between the various degrees of spontaneity.

Comparing the two experimental conditions (matched/mismatched), our ANOVA analysis revealed similar performance, for both speaking styles. This result has implications for both the quantity and the type of data necessary to train an automatic detection system. The mismatched conditions had a minimum of more than one hour of data (Kiel_L), reaching around 4 hours for Kiel_R and Kiel_A. However, both the conversational and the read matched data was comprised of less than 10 minutes of recordings. The fact that the matched condition gave a similar performance for the two speaking styles, despite using one order of magnitude less data than the mismatched condition, is encouraging for the annotation of larger quantities of data in less-resourced language varieties. It suggests that, when larger annotated datasets from a well-resourced variety are not available, one can rely of smaller amounts of annotated data in the target variety.

Here, we used three different algorithms for prosodic boundary detection, based on the same acoustic features. The conducted analyses showed that differences exist between the employed algorithms with the supervised system slightly outperforming the rule-based system, while being similar to the unsupervised approach. Comparing our results to those of previous approaches for boundary detection in German (cf. Table 7 for a summary), one can observe that the best boundary detection systems employing the four investigated cues (i.e., Kiel RB for read speech and GRASS NB for conversational speech) outperform most of them. Strom (1995) reported classification rates of between 33% and 67% for different levels of prosodic boundaries, in recordings of conversational dialogues. Similar detection rates of 56% were also found by Braunschweiler (2003), yet the author provides no breakdown of these values into the accent and boundary tones, respectively. Our algorithms reached a higher F-score for both conversational and read speech compared to those approaches. We obtained comparable results on conversa-

Table 7: Comparison of automatic prosodic boundary detection performance on German read and conversational speech, respectively.

speaking style	System	Best performance
Read	Braunschweiler (2003)	0.560[a]
	Soto et al. (2013)	0.910[b,c]
	Stehwien et al. (2020)	0.916[b]
	Proposed	0.824
Conversational	Strom (1995)	0.670
	Batliner et al. (2001)	0.758[b]
	Proposed	0.766

[a] Included also accent tone detection in the evaluation
[b] Used also word-level knowledge
[c] Employed also speaker identity information

tional speech to previous systems Batliner et al. (2001), although in that case also word-level information was known to the system and only the recall was reported for the evaluation of the detection task. Compared to the boundary tone classification performance presented by Soto et al. (2013), 91% on broadcast news, our best results on read speech are lower than that, but their system took a per-word decision, exploiting also word-level information and having knowledge of the identity of the speaker. Similar performance to Soto et al. (2013) on German read speech was reported also by Stehwien et al. (2020) when using acoustic features and word information.

Most of the systems presented here reached a good boundary detection performance, making them suitable candidates for being used in a semi-automatic boundary annotation process, as a first pass. Moreover, since the systems also exhibit a relatively high precision, the annotators correcting the boundary placed after the first (automatic) step will not require an extensive period of time for correcting the placed boundaries (and, thus, defeating the purpose of using an automated system - reducing the time required for annotation).

The proposed systems use only acoustic information to determine the prosodic boundaries, as this type of information is available for speech corpora which have been orthographically transcribed and force-aligned. However, other types of information, not considered here, are available and may be exploited successfully for boundary detection. Previous approaches used

knowledge extracted from other levels, not just that of the syllable, such as word or speaker information (Batliner et al., 2001; Soto et al., 2013; Stehwien et al., 2020). Some preliminary experiments, which used information on word boundaries to constrain the placement of prosodic boundaries showed that this simple strategy increases the precision of the systems by up to 7% and their F-score by up to 4%, on the conversational data. Further work will be conducted in this direction.

6. General discussion and conclusions

We studied here the acoustic cues to prosodic boundary marking in read and conversational speech of two varieties of German, with German data from the Kiel corpus and Austrian data from the GRASS corpus. For our analysis we employed several independent methods and compared their results: (1) mixed effects logistic regression, to determine the role of the investigated cues (pause, nucleus and onset duration, f0 reset) for marking boundaries, (2) Random Forest, to estimate the importance of these cues for the learning of boundaries, and (3) using them in three automatic boundary detection systems (rule-based, unsupervised and supervised).

Based on a total of more than 145k syllables (of which almost 24k preceding a prosodic boundary) extracted from the two corpora, we observed that pause duration is the strongest cue for prosodic boundaries and that f0 reset is the weakest. In both varieties, prosodic boundaries in conversational speech were characterized by shorter nucleus duration and tended to be produced less frequently with a pause. When comparing the conversational speech of the two varieties, we found that Northern German speakers tend to produce boundaries with longer nucleus duration and higher f0 reset than the Austrian speakers, and that Austrians tend to make longer pauses.

Despite the differences with respect to the excursion size of the individual features, our Random Forest analysis revealed that the weights given to the features were similar for conversational speech, in the two varieties. In read speech, however, the weighting of the features proved to be different between the varieties. Thus, from the point of view of feature ranking, the two varieties are more similar in conversational than in read speech. It remains to be determined by means of perception experiments, whether listeners are sensitive to the absolute values of the acoustic cues or to their relative weight/ranking.

The investigated acoustic cues were then tested in several automatic boundary detection systems, with the aim of facilitating the ongoing prosodic annotation process of the GRASS corpus. All three different approaches for boundary detection achieved an overall high performance, with a higher boundary detection rate for read speech (a maximum F-score of 82.4%) than for conversational speech (an F-score of 76.6%). Training on more data from the well-resourced variety did not outperform the same system trained on less data from the target variety, which indicates that the proposed method is a valuable approach for creating prosodic annotations in the absence of large quantities of available manual labels.

Each approach employed here has revealed different aspects of the boundary marking process in the two German varieties and across speaking styles and, overall, their findings seem to support each other. With respect to speaking style, our feature importance analysis showed that differences exist between the two conditions in the weight given to the pause and the onset cue for marking boundaries. The same two cues exhibited the highest effects in the statistical models fitted on the two corpora and exhibited also significant interactions with speaking style. Moreover, these differences in the acoustic marking of the boundaries were reflected in the automatic detection process, with boundaries in read speech being more reliably detected than in conversational speech. However, when looking at the differences between varieties, the three analyses agree only partly. Although the fitted regression models showed a number of significant differences between the two varieties, both for conversational as well as for read speech (Schuppler and Ludusan, 2020), it seems that these differences did not have an effect on the boundary detection process. The results of the feature importance analysis (showing similar weighs given to the acoustic cues in the two varieties) are in line with those of the automatic detection task, indicating similarities between the two varieties.

The results of our statistical and feature importance analyses agree also with the findings of other studies on prosodic boundary detection (e.g., Batliner et al. 2001). Using two different approaches for ranking the importance of the acoustic cues (linear discriminant analysis and decision trees), Batliner et al. (2001) reported that the best discriminating features for prosodic boundary marking in conversational German were related to duration, speech intensity and pauses, while f0 information was found to be less important. While we did not consider any intensity-based measures in this study, it has been previously shown that intensity levels may be used to discriminate

prosodic boundaries in various languages (Ludusan and Dupoux, 2015) and it has been employed also in other works on automatic boundary detection in German (e.g., Strom 1995). A possible future research direction may involve the inclusion of intensity-related cues in acoustic analyses of prosodic boundaries.

Acknowledgements

The work done by Barbara Schuppler was funded by an Elise Richter grant (V638 N33) from the Austrian Science Fund. It was also supported by Grant No. P 32700 from the Austrian Science Fund. The authors would like to thank the annotators David Ertl, Anneliese Kelterer and Katerina Petrevska for their efforts, Simon Wasserfall for his support with the automatic segmentation of the GRASS corpus, and two anonymous reviewers for their valuable and constructive feedback.

References

Ananthakrishnan, S., Narayanan, S. S., 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Transactions on Audio Speech and Language Processing 16 (1), 216–228.

Apel, J., Neubarth, F., Pirker, H., Trost, H., 2004. Have a break! Modelling pauses in German speech. In: Proceedings of KONVENS. pp. 5–12.

Baayen, R. H., 2008. Analyzing linguistic data. A practical introduction to statistics using R. Cambridge University Press.

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

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 2001. Boiling down prosody for the classification of boundaries and accents in German and English. In: Proceedings of EUROSPEECH. pp. 2781–2784.

Beckman, M., Edwards, J., 1990. Lengthenings and shortenings and the nature of prosodic constituency. In: Kingston, J., Beckman, M. (Eds.), Papers on Laboratory Phonology 1: Between the grammar and physics of speech. Cambridge University Press, Cambridge, pp. 152–178.

Biadsy, F., Hirschberg, J., 2009. Using prosody and phonotactics in Arabic dialect identification. In: Proceedings of INTERSPEECH. pp. 208–211.

Braun, B., Einfeldt, M., Esposito, G., Dehé, N., 2020. The prosodic realization of rhetorical and infromation-seeking questions in German spontaneous speech. In: Proceedings of Speech Prosody. pp. 342–346.

Braunschweiler, N., 2003. ProsAlign - The Automatic Prosodic Aligner. In: Proceedings of ICPhS. pp. 3093–3096.

Cho, T., Keating, P., 2009. Effects of initial position versus prominence in English. Journal of Phonetics 37, 466–485.

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: Proceedings of INTERSPEECH. pp. 3936–3940.

Christophe, A., Gout, A., Peperkamp, S., Morgan, J., 2003. Discovering words in the continuous speech stream: The role of prosody. Journal of Phonetics 31 (3-4), 585–598.

Church, R., Bernhardt, B., Pichora-Fuller, K., Shi, R., 2005. Infant-directed speech: Final syllable lengthening and rate of speech. Canadian Acoustics 33 (4), 13–19.

Clements, G. N., 1990. The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M. (Eds.), Papers on Laboratory Phonology 1: Between the grammar and physics of speech. Cambridge University Press, Cambridge, pp. 283–333.

Cutler, A., Dahan, D., Van Donselaar, W., 1997. Prosody in the comprehension of spoken language: A literature review. Language and Speech 40 (2), 141–201.

De Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America 111 (4), 1917–1930.

De Pijper, J. R., Sanderman, A. A., 1994. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. The Journal of the Acoustical Society of America 96 (4), 2037–2047.

El Zarka, D., Schuppler, B., Cangemi, F., 2019. Acoustic cues to topic and narrow focus in Egyptian Arabic. In: Proceedings of INTERSPEECH. pp. 1771–1775.

El Zarka, D., Schuppler, B., Lozo, C., Eibler, W., Wurzwallner, P., 2017. Acoustic correlates of stress and accent in Standard Austrian German. In: Moosmüller, S., Schmid, C., Sellner, M. (Eds.), Phonetik in und über Österreich, Veröffentlichungen zur Linguistk und Kommunikationsforschung: 31. ÖAW Austrian Academy of Sciences Press, Vienna, pp. 15–44.

Feizollahi, Z., Soukoup, B., 2011. The role of intonation in Austrian listeners' perceptions of standard-dialect shifting. In: Gregersen, F., Parrott, J. K., Quist, P. (Eds.), Language Variation – European Perspectives III. John Benjamins Publishing Company, Amsterdam, pp. 31–42.

Fletcher, J., 2010. The prosody of speech: Timing and rhythm. Wiley Online Library, Ch. 15, pp. 523–602.

Fougeron, C., Keating, P. A., 1997. Articulatory strengthening at edges of prosodic domains. The Journal of the Acoustical Society of America 101 (6), 3728–3740.

Fuchs, S., Krivokapic, J., Jannedy, S., 2010. Prosodic boundaries in German: Final lengthening in spontaneous speech. Journal of the Acoustical Society of America 127 (3), 1851.

Gubian, M., Torreira, F., Strik, H., Boves, L., 2009. Functional data analysis as a tool for analyzing speech dynamics. A case study on the French word c'était. In: Proceedings of INTERSPEECH. pp. 2199–2202.

Hagmüller, M., 2001. Recognition of regional variants of German using prosodic features. Master's Thesis, Graz University of Technology.

Holzgrefe-Lang, J., Wellmann, C., Petrone, C., Räling, R., Truckenbrodt, H., Höhle, B., Wartenburger, I., 2016. How pitch change and final lengthening cue boundary perception in German: converging evidence from ERPs and prosodic judgements. Language, Cognition and Neuroscience 31 (7), 904–920.

Kim, J., 2019. Individual differences in the production of prosodic boundaries in American English. In: Proceedings of ICPhS. pp. 1024–1028.

Kim, S.-E., Tilsen, S., 2020. Speech rate and syntactically conditioned influences on prosodic boundaries. In: Proceedings of Speech Prosody. pp. 434–438.

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

Kohler, K. J., 2006. Paradigms in experimental prosodic analysis: From measurement to function. In: Methods in empirical prosody research. De Gruyter, pp. 123–152.

Kohler, K. J., Peters, B., Scheffers, M., 2017. The Kiel Corpus of Spoken German—Read and Spontaneous Speech. New Edition, revised and enlarged. Available at http://www.isfas.uni-kiel.de/de/linguistik/forschung/kiel-corpus/.

Levshina, N., 2015. How to do Linguistics with R. Data exploration and statistical analysis. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Leykum, H., 2019. Acoustic characteristics of verbal irony in Standard Austrian German. In: Proceedings of ICPHS. pp. 3398–3402.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., Dupoux, E., 2016. Learnability of prosodic boundaries: Is infant-directed speech easier? The Journal of the Acoustical Society of America 140 (2), 1239–1250.

Ludusan, B., Dupoux, E., 2014. Towards low-resource prosodic boundary detection. In: Proceedings of SLTU. pp. 231–237.

Ludusan, B., Dupoux, E., 2015. A multilingual study on intensity as a cue for marking prosodic boundaries. In: Proceedings of ICPhS. p. 982.

Ludusan, B., Wagner, P., Włodarczak, M., 2021. Cue Interaction in the Perception of Prosodic Prominence: The Role of Voice Quality. In: Proceedings of INTERSPEECH. pp. 1006–1010.

Luthern, E., Clopper, C. G., 2015. Variation in glottalization at prosodic boundaries in clear and plain lab speech. In: Proceedings of ICPhS. pp. 352–355.

Männel, C., Friederici, A. D., 2016. Neural correlates of prosodic boundary perception in German preschoolers: If pause is present, pitch can go. Brain Research 1632, 27–33.

Markó, A., Kohári, A., 2015. Glottalization and timing at utterance final position in Hungarian: Reading aloud vs. spontaneous speech. In: Proceedings of ICPhS. p. 722.

Megyesi, B., Gustafson-Čapková, S., 2002. Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish. In: Proceedings of INTERSPEECH. pp. 2153–2156.

Mo, Y., Cole, J., 2010. Perception of prosodic boundaries in spontaneous speech with and without silent pauses. The Journal of the Acoustical Society of America 127 (3), 1956.

Moosmüller, S., 2015. The interaction of prosody and phonotactics: Resyllabification in three varieties of German. Italian Journal of Linguistics 27 (1), 111–132.

Moosmüller, S., Brandstätter, J., 2014. Phonotactic information in the temporal organization of Standard Austrian German and the Viennese dialect. Language Sciences 46, 84–95.

Morrill, T., Baese-Berk, M., Bradlow, A., 2016. Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of English. In: Proceedings of Speech Prosody. pp. 1119–1123.

Neubarth, F., Alter, K., Pirker, H., Rieder, E., Trost, H., 2000. The Vienna prosodic speech corpus: Purpose, content and encoding. In: Proceedings of KONVENS. pp. 191–195.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peters, B., 2003. Multiple cues for phonetic phrase boundaries in German spontaneous speech. Proceedings of ICPhS, 1795–1798.

Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., Höhle, B., 2017. Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. Journal of Phonetics 61, 71–92.

Pirker, H., Neubarth, F., 2003. Some questions and answers on the prosodic correlates of information structure. In: Proceedings of ICPhS. pp. 1807–1810.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Sadat-Tehrani, N., 2017. Intonation of Persian declaratives: Read vs. spontaneous speech. Questions and Answers in Linguistics 4 (1), 21–43.

Schleef, E., 2003. Prosody and narrative structure in varieties of Low German and Alemannic. Journal of Germanic Linguistics 15 (4), 325–357.

Schuppler, B., Adda-Decker, M., Morales-Cordovilla, J. A., 2014a. Pronunciation variation in read and conversational Austrian German. In: Proceedings of INTERSPEECH. pp. 1453–1457.

Schuppler, B., Grill, S., Menrath, A., Morales-Cordovilla, J. A., 2014b. Automatic phonetic transcription in two steps: forced alignment and burst detection. In: Besacier, L., Dediu, A., Martín-Vide, C. (Eds.), Statistical Language and Speech Processing. SLSP 2014. Lecture Notes in Artificial Intelligence. Vol. 8791. Springer, pp. 132–143.

Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H., 2014c. GRASS: the Graz corpus of Read And Spontaneous Speech. In: Proceedings 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., Kelterer, A., 2021. Developing an annotation system for communicative functions for a cross-layer ASR system. In: Proceedings of the Integrating Perspectives on Discourse Annotation Workshop. p. 3.

Schuppler, B., Ludusan, B., 2020. An analysis of prosodic boundary detection in German and Austrian German read speech. In: Proceedings of Speech Prosody. pp. 990–994.

Schwab, S., Avanzi, M., 2015. Regional variation and articulation rate in French. Journal of Phonetics 48, 96–105.

Schweitzer, A., Lewandowski, N., Duran, D., Dogil, G., 2015. Attention, please! Expanding the GECO database. In: Proceedings of ICPhS. p. 620.

Sertling Miller, J., 2007. Swiss French prosody: intonation, rate and speaking style in the Vaud Canton. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Siddins, J., Mennen, I., 2019. Pitch accent realisation in Austrian German. In: Proceedings of ICPHS. pp. 2846–2850.

Silverman, K., Blaauw, E., Spitz, J., Pitirelli, J. F., 1992. A prosodic comparison of spontaneous speech and read speech. In: Proceedings of ICSLP. pp. 1299–1302.

Simon, A. C., Christodoulides, G., 2016. Perception of prosodic boundaries by naïve listeners in French. In: Proceedings of Speech Prosody. pp. 1158–1162.

Skarnitzl, R., Machac, P., 2011. Principles of phonetic segmentation. Phonetica 68, 198–199.

Soto, V., Cooper, E., Rosenberg, A., Hirschberg, J., 2013. Cross-language phrase boundary detection. In: Proceedings of ICASSP. pp. 8460–8464.

Soukup, B., December 2007. The strategic use of Austrian dialect in interaction: A sociolinguistic study of contextualization, speech perception and language attitudes. Ph.D. thesis, Georgetown University.

Stehwien, S., Schweitzer, A., Vu, N. T., 2020. Acoustic and temporal representations in convolutional neural network models of prosodic events. Speech Communication 125, 128–141.

Strom, V., 1995. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In: Proceedings of EUROSPEECH. pp. 2039–2042.

Swerts, M., 1997. Prosodic features at discourse boundaries of different strength. The Journal of the Acoustical Society of America 101 (1), 514–521.

Swerts, M., Strangert, E., Heldner, M., 1996. F/sub 0/ declination in read-aloud and spontaneous speech. In: Proceeding of ICSLP. pp. 1501–1504.

Trouvain, J., Grice, M., 1999. The effect of tempo on prosodic structure. In: Proceedigns of ICPhS. pp. 1067–1070.

Ulbrich, C., 2006. Prosodic phrasing in three German standard varieties. In: Proceedings of 29th Annual Penn. Linguistics Colloquium. pp. 361–373.

Vaissière, J., 2005. Perception of intonation. Wiley Online Library, Ch. 10, pp. 236–263.

Velázquez, E., 2010. Acoustic comparative study of Spanish prosody. Mexico City vs. Madrid. In: Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology. pp. 83–90.

Verhoeven, J., De Pauw, G., Kloots, H., 2004. Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. Language and Speech 47 (3), 297–308.

Volín, J., Weingartová, L., Niebuhr, O., 2014. Between recognition and resignation – the prosodic forms and communicative functions of the Czech confirmation tag "jasně". In: Proceedings of Speech Prosody. pp. 115–119.

Wang, X., Li, A., Yuan, C., 2008. A preliminary study on silent pauses in Mandarin speech. In: Proceedings of Speech Prosody. pp. 673–676.

Ward, N., 2019. Prosodic Patterns in English Conversation. Cambridge University Press.

Wasserfall, S., 2020. Automatic speech segmentation using Kaldi. Master's Thesis, Graz University of Technology.

White, L., Wiget, L., Rauch, O., Mattys, S. L., 2010. Segmentation cues in spontaneous and read speech. In: Proceedings of Speech Prosody. p. 218.

Yan, Q., Vaseghi, S., 2010. Modeling and synthesis of English regional accents with pitch and duration correlates. Computer Speech & Language 24 (4), 711–725.

Yang, Y., Wang, B., 2002. Acoustic correlates of hierarchical prosodic boundary in Mandarin. In: Proceedings of Speech Prosody. pp. 707–710.

Yoon, T.-J., Cole, J., Hasegawa-Johnson, M., 2007. On the edge: Acoustic cues to layered prosodic domains. In: Proceedings of ICPhS. pp. 1264–1267.