Journal of Phonetics 117 (2026) 101503

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

# Temporal alignment of prosody and gesture in focus: unimodal and multimodal hyperarticulation

Alina Gregori *, Frank Kügler

*Goethe University Frankfurt, Norbert-Wollheim-Platz 1, 60629 Frankfurt am Main, Germany*

ARTICLE INFO

*Article history:*
Received 6 June 2025
Received in revised form 11 May 2026
Accepted 11 May 2026
Available online xxxx

*Keywords:*
Co-speech gestures
f0-slopes
Multimodality
Gesture-prosody coordination
Temporal alignment
Hyperarticulation
Focus

ABSTRACT

This study investigates how the temporal alignment of prosody and gesture is affected by focus in German, examining both unimodal (prosodic) and multimodal (gesture-prosody) hyperarticulation. Specifically, we address two research questions: (i) does the realization of German falling pitch accents vary as a function of focus (versus background), and (ii) does gesture–prosody alignment vary as a function of focus? Drawing on Hypo- and Hyperarticulation (H&H) theory which assumes that speakers will adapt the articulatory precision of speech to signal the importance of specific information to a listener, we hypothesize that focused constituents show tighter unimodal prosodic and multimodal temporal alignment between gesture and prosody than background constituents.

We conducted a corpus study on German data from the SaGA corpus, running two analyses: a unimodal prosodic analysis of the slopes of falling accents (n = 345) and a multimodal analysis of the temporal alignment of gesture apexes and pitch accents (n = 1801). Both analyses compare cues in focus and in background. Results on the falling accents show that steeper slopes are produced in (contrastive) focus compared to in background contexts, indicating a tighter alignment of the f0-fall with the focused constituent. Similarly, results on the temporal alignment show that gesture apexes are more closely aligned with accents in information and contrastive focus than in background. This study not only confirms prosodic hyperarticulation in focus marking in German, it allows H&H-theory to be extended to an additional modality by interpreting the closer apex-accent alignment in focus as multimodal hyperarticulation.



## 1. Introduction

In communication, speakers transfer information to their conversation partner, and they adapt their speech to ensure that all relevant information of an utterance can be processed by the interlocutor. This communicative goal is captured in the Hypo- and Hyperarticulation theory (H&H-theory, Lindblom, 1990), which states that speakers orient their speech towards their listener by producing a more precisely articulated acoustic signal when necessary to convey all important information (hyperarticulation) and reducing their articulatory effort where possible (hypoarticulation), for instance when the context facilitates the processing of information. Hence, speech varies on a continuum from hypo- to hyperarticulation.

Hyperarticulated elements are often perceived as more prominent than other elements, but these concepts are orthogonal to each other. Generally, the concept of prominence in linguistics is relational to its context. In other words, “a prominent element organizes its environment, providing a structure for the context in which it appears [. . .]” (Grice & Kügler, 2021, p. 253). Prominence is of interest in different linguistic domains and is situated at their interface, as prominence often aligns across domains, for instance in prosody and pragmatics. The relation of prosody and pragmatics is at the core of interest for the present study. We use focus as the pragmatic domain of prominence in this article, which is characterized by the choice of a referent from a set of alternatives. Focus constituents have increased pragmatic prominence compared to their context, which is usually reflected in prosodic marking: Focused constituents have been reported to be marked by increasing prosodic prominence in West Germanic languages (Baumann et al., 2006; Féry & Kügler, 2008; Kügler & Calhoun,

* Corresponding author.
*E-mail addresses:* gregori@lingua.uni-frankfurt.de (A. Gregori), kuegler@em.uni-frankfurt.de (F. Kügler).

https://doi.org/10.1016/j.wocn.2026.101503
0095-4470/

2020). More specifically, prosodic prominence refers to the composition and strength of a set of acoustic features (like f0, duration or intensity) which make a referent stand out acoustically from other constituents. Some characteristics of the prominence dimension are similar to the conceptualization of H&H-theory, but they need to be clearly disentangled from each other, since they represent orthogonal characteristics of prosody.

Communication is inherently multimodal, as it does not solely rely on the acoustic channel (which would be unimodal), but also on other modes of communication, like the visual modality (McNeill, 1992; Perniss, 2018; Wagner et al., 2014). Visual aspects of communication, like speech-accompanying gestures, have become a topic of formal linguistics rather recently (cf. an overview in Gregori et al., 2023). For instance, in the last decades, researchers have started to investigate the coordination and alignment of prosody and gestures as an integral multimodal part of language (McNeill, 1992; Shattuck-Hufnagel & Ren, 2018; Wagner et al., 2014; Prieto et al., 2024). Research on the temporal alignment of prosody and gesture is limited, especially concerning the variation along different linguistic contexts like focus (see however first studies on multimodal focus marking, Türk & Calhoun, 2024; Im & Baumann, 2020). The present study adds to this line of research, investigating how different pragmatic contexts influence the precision of articulation in the prosodic and visual channels of communication. We explore variation in multimodal alignment in focus and background, investigating how prosodic and gestural landmarks (reference points of each modality) are integrated temporally in the marking of focus. We examine whether visual cues contribute to the expression of focus by exploring H&H-theory in a multimodal setting.

## 2. Background

### 2.1. Focus

Conceiving focus as a cognitive domain of information structure (Chafe, 1976; Lambrecht, 2000), we follow Krifka in defining focus as "indicat[ing] the presence of alternatives that are relevant for the interpretation of linguistic expressions" (Krifka, 2008, p. 247). Focus thus signals the importance of a referent and its selection over other possible candidates in a discourse.

The discourse context is explicitly or implicitly (e.g., in the case of Question under Discussion, Roberts, 2012) relevant when assessing focus. Different focus types have been proposed, and in this study, we consider information focus (which refers to new or the most important sentence information, Götze et al., 2007) and contrastive focus (which "evokes a notion of contrast to another element", Götze et al., 2007, p. 178). Contrastive focus has been claimed to bear more prominence than information focus, due to the explicit contrast to other constituents (Zimmermann, 2008; Repp, 2016, Cruschina; 2021). Sentence constituents that are not in focus are background constituents. Generally, the scope of focus can be narrow (from only one to a few words) or broad (verb phrase or sentence focus), depending on the discourse context (Ladd, 1980). Prosodic effects of scope size have been addressed in previous studies (e.g., Baumann et al., 2006; Hanssen et al., 2008) but are not considered in this study. Focus is known to be marked by prosody in West Germanic languages (Kügler & Calhoun, 2020), which is discussed in detail in the next section.

### 2.2. Prosodic marking of focus in German

According to Kügler & Calhoun (2020), a focused word is the most prosodically prominent one in an utterance. Phonologically, this prominence is expressed by a pitch accent. Pitch accents are prosodic landmarks characterized mainly by fundamental frequency (henceforth f0) and can be identified based on local f0-maxima, f0-minima and f0-contours (Ladd, 2008). These features enable the categorization of a finite number of pitch accent types, which signal differences in prominence (Baumann et al., 2006; Baumann & Röhr, 2015). Pitch accents are associated with stressed syllables in Germanic languages. Specifically relevant for the present study are falling pitch accents, as it has been found for Dutch that they show variation in their falling contour related to the pragmatic context in which they are produced (Hanssen et al., 2008). Falling pitch accents are characterized in GToBI, the pitch accent annotation system for German intonation, as an f0-minimum on an accented syllable which is immediately preceded by an associated f0-maximum (Grice et al., 2005).

Differences in prosodic prominence have been shown to be associated with pragmatic prominence in West Germanic languages (Kügler & Calhoun, 2020; Baumann & Röhr, 2015). For example, high or falling pitch accents tend to be associated with focused elements rather than with background elements (see Féry, 1993; Féry & Kügler, 2008; Grice et al., 2009; Kügler & Calhoun, 2020 for an overview). Based on their general ability to "convey 'post-lexical' or sentence-level pragmatic meaning in a linguistically structured way" (Ladd, 2008, p. 4), we take pitch accents to be indicators of prosodic prominence in stress accent languages (Beckman, 2010; henceforth intonation languages) like German.

While pitch accents are associated with stressed syllables in traditional phonological theories, studies on tonal alignment have pointed out the relevance of the timing of f0-maxima and f0-minima for communicative and cognitive phenomena (Atterer & Ladd, 2004; Kohler, 2005; Ladd & Morton, 1997; Niebuhr & Kohler, 2004; Xu & Xu, 2005). For instance, Hanssen et al. (2008) found for Dutch that the slopes of falling f0-contours vary as a function of focus, such that steeper slopes are found in narrow compared to broad focus contexts. Later (compared to earlier) pitch peaks have often been described as perceived to be associated with more contrastive focus types by the aforementioned studies. Further studies on phonetic aspects of focused constituents in West Germanic languages (e.g., Baumann et al., 2006, 2007; Sityaev & House, 2003; Féry & Kügler, 2008; Kügler & Gollrad, 2015; Kohler, 2005; Elsner, 2000) have found that f0-extrema are produced later and f0-peaks are higher in focus, and that pitch accents employ a higher f0-range and a steeper slope in focus.

### 2.3. Hypo- and hyperarticulation theory and the effort code

A relevant theory for assessing the variation in speech is hypo- and hyperarticulation theory, which was proposed by

Lindblom (1990, H&H-theory). H&H-theory claims that speakers adapt their articulation in terms of clarity to conform to the conversational needs of their interlocutor(s) and, what is important to our study, to signal the importance of information: Words or phrases are produced with maximum acoustic information (hyperarticulation) when their contribution cannot be derived or supplemented from the linguistic context. Hyperarticulated elements are produced more clearly, precisely and generally enhanced, with the goal of differentiating them from other elements. As a purely phonetic example, when producing loud speech, speakers not only increase the intensity of sounds, they also increase sound duration and vocalic duration, which is necessary to fulfill physiological aspects of loud speech, like larger jaw openings (Lindblom, 1990; after Schulman, 1989). In a similar vein, articulatory effort can be reduced (hypoarticulation), making words or phrases less distinct or weaker, when the linguistic context already facilitates comprehension. In the example of loud speech, hypoarticulation can be observed in the surrounding acoustic elements, which are not only reduced in intensity, but also in duration. This is done to compensate for the loud elements, but also because this information has less relevance for the utterance (Lindblom, 1990; after Schulman, 1989). Hypoarticulation has a lower limit on how reduced a sound can be, requiring reduced elements to still be distinguishable from lexical competitors. According to H&H-theory, utterances are situated on a continuum of hyper- and hypoarticulation. Hyperarticulation is thus linked to the addressee (e.g., Picheny et al., 1985; Kuhl et al., 1997) and to the speaking style of the speaker (Moon & Lindblom, 1994).

Fig. 1 shows a schematic illustration of our interpretation of hypo- and hyperarticulation. For a given acoustic cue, the variation in articulation manifests itself as more extreme values in a tighter distribution around a mean in hyperarticulation, whereas it appears as less extreme values in a more spread-out distribution around a mean for hypoarticulation. Note that the scales in Fig. 1 depend on the type of measure expressing either higher or lower mean or SD values depending on which measure indicates higher precision. The graph in Fig. 1 shows both a higher mean and a smaller standard deviation in the enhanced data, but just one of these measures is sufficient to signal hyperarticulation.

Evidence for the correlation of prosodic prominence and hyperarticulation was presented by De Jong (1995), who discussed that stress in English, induced by linguistic prominence, increases distinctions in sound articulation. Acoustically prominent elements are frequently hyperarticulated and vice versa, but not always. We consider hyperarticulation and prominence to be orthogonal to each other. For instance, as prominence is a relational concept, an element can also stand out, and therefore be prominent, by being articulatorily reduced and thus hypo- instead of hyperarticulated.

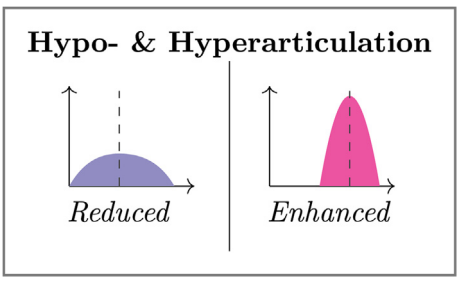


**Fig. 1.** Schematic illustration of the concept of hypo- and hyperarticulation. The left distribution shows reduced data, with a lower mean value for a given acoustic feature (dashed line) and a higher standard deviation (wider spread distribution of values). The right distribution shows enhanced data, with a higher mean value for a given acoustic cue (dashed line) and a smaller standard deviation (narrower distribution of values).

H&H-theory, and especially hyperarticulation, is compatible with the *Effort Code* (Gussenhoven, 2002, 2004), which states that "the amount of energy expended on speech production can be varied: putting in more effort will not just lead to more precise articulatory movements, but also to more canonical and more numerous pitch movements" (Gussenhoven, 2002, p. 1). Note that there is a difference between physical effort and communicative effort, which do not necessarily need to align with one another. Hyperarticulation operates on a communicative effort level (as hyperarticulated parts can be 'easier' to produce), while the Effort Code covers both aspects. The Effort Code not only supports the idea that speakers adapt their productions to vary articulatory precision by varying their level of effort, but Gussenhoven also transfers the concept of the Effort Code to intonation in production and meaning. As an empirical example, Hanssen et al. (2008) conducted an experimental study on Dutch intonation contours. In their lab study, participants produced one lexically identical sentence in different focus contexts eliciting narrow information focus and corrective focus, or broad focus. This led to the production of different f0-contours as a marker of focus: target words were produced with a steeper f0-fall on the accented syllable and the following syllable in narrow focus than in broad focus constituents. The difference between the falling contours in the study on Dutch was expressed in terms of slope calculations done by measuring the difference between f0-height at the start of the f0-fall up to the first post-stressed syllable. In their case, narrow focus included information focus and corrective focus contexts, which showed similar contours. The steeper f0-falls in focus can be interpreted as a prosodic instantiation of hyperarticulation within different focus contexts, indicating a more precise intonational alignment of prosodic landmarks with segments and syllables. This instantiation of H&H theory shows a clear indication of acoustic focus marking contrasts and in addition shows that focus is a source of articulatory variation. In German, however, prosodic hyperarticulation has not been addressed in similar terms. The present study therefore aims to investigate whether similar patterns of f0-slopes of falling accents are found for German as they were found for Dutch.

### 2.4. Gestures and their role in focus marking

In the visual modality, as part of multimodal communication, gestures have been defined as "visible bodily action" (Kendon, 2004, p. 1) that makes a relevant contribution to communication. Co-speech gestures are those gestures that accompany speech (in contrast to pro-speech gestures, which replace speech; cf. Ladewig, 2011; Schlenker, 2019). Gestures reinforce or complement speech, for instance by specifying something that has been said. If a speaker says, *'My dog picked up a stick at the beach.'* while accompanying '*a stick*' with a gesture, the gesture could specify the shape or size of that stick,

resulting in a more informative utterance. In Fig. 2, the gesture in picture a) indicates that the stick would be rather small, whereas the gesture in picture b) would indicate a larger stick. Therefore, in both cases, the gesture adds information about the size of the object that is mentioned in speech and produces a difference in meaning between the two utterances.

Originally, McNeill (1992) first categorized gestures into different types based on their referential (semantic) connection to speech. Iconic and metaphoric gestures depict a referent in speech directly or abstractly, deictic gestures provide spatial information and beat gestures do not have a semantic connection to speech. The idea of discrete semantic gestural categories has been revised and superseded by a dimensional system by McNeill himself (in 2006) due to the original suggestion being restrictive in not allowing gestures to have multiple semantic and non-semantic functions. More recent gesture classification systems assume multiple gesture dimensions in which a gesture receives labels of features in multiple dimensions (e.g., M3D; Rohrer et al., 2025). For instance, the M3D annotation system allows the superimposition of 'beat-like' components onto any type of gesture, calling gestures without a semantic connection to speech 'non-referential'. We adopt this term for this article.

When gesturing, speakers use their hands, torso, head or eyebrows, potentially in combination. For manual gestures, a hierarchical structure has been proposed, meaning that they comprise differently sized components that are nested into each other (Kendon, 1980). Kendon (1980) claims that the gestural stroke is the only obligatory component of a gesture, representing the prominent core movement and carrying the main communicative contribution. The apex is the peak of a gestural stroke; it is a point in time and is defined as the kinematic goal of the stroke (Rohrer et al., 2025; after Loehr, 2007, p. 189). It is an end or turning point of a gestural stroke movement or a point of no velocity in the movement and is identified as the maximum extension of the involved body part (see Rohrer et al., 2025; Yasinnik et al., 2004). The apex is the investigated constituent in this article. Other, optional, gesture phases can enclose a gesture stroke: preparation, hold and recovery. Together with the stroke, they form a complete gesture phrase. Adjacent gesture phrases then form bigger gesture units. Fig. 3 illustrates the hierarchical composition of gestures.

The characteristics of a gesture apex make it a suitable component to relate to f0-extrema of pitch accents since both are landmarks that represent points in time, making them discrete gestural and prosodic targets.

A number of studies in the field have investigated the gestural marking of information structure. These studies indicated that the distribution of gestures varies as a function of information structure. Concerning the information status of referents, Im & Baumann (2020) found more gestures on new and accessible referents than on given referents in English. Similar results have been found for German, English and French by Debreslioska & Gullberg (2020), Rohrer (2022) and Baills & Baumann (2025, head gestures). In a broader pragmatic sense, gestures have been reported to be enhanced in their kinematic characteristics (size, complexity of the movement, precision) when accompanying a new discourse referent (Gerwing & Bavelas, 2005). In terms of focus marking, Türk (2020) found that in Turkish, gestures associate more frequently with focus (and topic) than with background constituents. Esteve-Gibert et al. (2022) found an increased number of non-manual gestures on contrastively focused constituents compared to broad focus constituents in French-speaking children, and Coego et al. (2025) found a similar trend for manual and non-manual gestures in Catalan-speaking children. Ebert and colleagues (2011) showed that gestures are relevant in focus marking, contributing to the disambiguation and interpretation of sentences. In general, gestures have been found to associate more with pragmatically stronger discourse referents, resembling an effect of prosody with strengthened prosodic cues on pragmatically stronger discourse referents.

### 2.5. Focus marking in a multimodal setting

It is widely accepted that communication is multimodal, and that gestures and speech coordinate with each other. Regarding this coordination, McNeill (1992) proposed three gesture-speech synchrony rules. In addition to the semantic and the pragmatic synchrony rule, which state that the semantic information and pragmatic function of speech and gestures must match or complement each other (see also McNeill & Duncan, 2000), McNeill established the *phonological synchrony rule*. This rule concerns the integration of prosody and gesture, and it states that "the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech" (McNeill, 1992, p. 26; based on Kendon, 1980). This means that gestures and speech have a close temporal relation in the sense that gestural movement (specifically a stroke phase) associates with prominent syllables, making it reasonable to investigate whether this relation varies between different contexts.

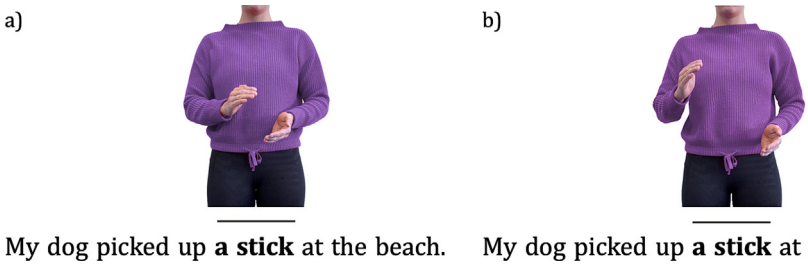


**Fig. 2.** Example of a co-speech gesture providing complementary information (about the size of an object) to a spoken utterance. The utterance in a) denotes a smaller size of the object, the utterance in b) denotes a bigger size of the object. The line above the phrase 'a stick' and the bold font indicate the placement of the gesture.

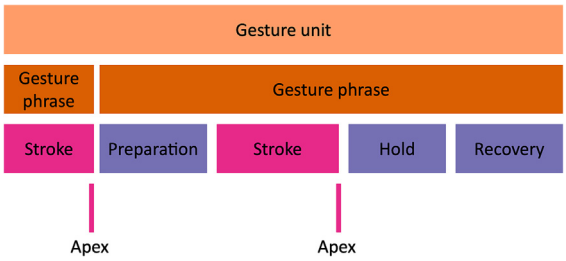


**Fig. 3.** The hierarchy of the gestural components following Kendon (1980) in an exemplary gesture unit.

Following McNeill's claims, the integration of gesture and speech has been empirically investigated from different methodological perspectives, allowing the concrete prosodic and gestural landmarks that associate with each other to be specified. Concerning the temporal alignment of gestures with prosody, it has been generally confirmed that gestures and prosodic landmarks align on the level of gesture stroke and syllable (Shattuck-Hufnagel & Ren, 2018; Yasinnik et al., 2004). Several studies found that gesture apexes and pitch accents are temporally associated with each other, in different communicative settings (Shattuck-Hufnagel et al., 2007; Leonard & Cummins, 2011; Loehr, 2012; Esteve-Gibert & Prieto, 2013; Kügler & Gregori, 2023). For example, Loehr (2012) found synchronization of pitch accents and gesture apexes, and a looser integration of gesture phrases with intermediate phonological phrases, considering phrase alignment. Similarly, Türk & Calhoun (2024) conducted an alignment study in Turkish, addressing prosodic and gestural phrase overlaps and on- and offset alignment. They found coordination of phrase on- and offsets, which changed in different communicative contexts. These studies indicate that gestural landmarks coordinate temporally with prosodic landmarks, at least in intonation languages (but see Fung & Mok, 2018; Rohrer et al., 2024 for other prosodic anchors in prosodically different languages). Krivokapić et al. (2017), Pouw and Dixon (2019) and Pouw et al. (2021) found gesture-prosody synchronization from a signal-based perspective using articulatory and kinematic data.

The three-way interaction between prosody, gestures and focus is less systematically explored. However, as it has been shown that both prosody (Kügler & Calhoun, 2020) and gestures (Ebert et al., 2011) show different behavior in focus and background contexts, it is evident that the interaction of prosody and gesture may vary as a function of focus. To get a more comprehensive picture of prosody-gesture interaction and multimodal focus marking, it is crucial to investigate the alignment of prosody and gesture in different focus contexts. Initial studies have gained insights into this interaction, agreeing that increasing pragmatic prominence (here: focus) is marked by increasing multimodal cues, concerning cues like gesture rates, temporal alignment, and the variation in gestural articulators and referentiality. For example, Im & Baumann (2020) showed a parallel increase in pitch accentuation and gesture rate in contrastive focus contexts compared to non-contrastive contexts for English. Gregori, Sánchez-Ramón et al. (2024) found by conducting a semi-spontaneous production study in Catalan and German that both prosodic and gestural prominence increase across focus types, indicating stronger multimodal prominence with a stronger notion of pragmatic contrast. A similar trend was found by Esteve-Gibert et al. (2022) and by Coego et al. (2025) with French- and Catalan-speaking children, respectively. Less is known about temporal alignment changing in focus, but Kügler & Gregori (2023) ran a pilot on German iconic gestures that suggests that both non-referential and iconic gestures display a narrower alignment in focus than in background, but without distinguishing focus types. Türk & Calhoun (2024) found temporal alignment changes in different focus and topic contexts in Turkish. In line with this, Lehečková et al., (2022) found alignment of gesture strokes with foci and topic in Czech, and earlier gesture strokes compared to prosodic peaks in contrastive focus over information focus.

These findings on prosody, gesture and focus indicate that multimodal prominence is shaped by the coordination of prosody and gesture to increase a general impression of prominence. This is conceptualized in the framework of the Cumulative-cue hypothesis (Ambrazaitis & House, 2022, 2023), which suggests that prosodic and gestural cues tend to increase cumulatively with increasing pragmatic prominence, which is the observed trend in the empirical studies introduced. The parallelism of the modalities signals an increase in multimodal prominence, which is realized both prosodically and gesturally. This can be seen in gesture and accent occurrence, the number of articulators involved or in acoustic and kinematic features. Thus, the core idea of the cumulative-cue hypothesis is that speakers make use of multiple modalities simultaneously in terms of cue strength to convey greater prominence, which is expressed in different coordinative dimensions. While coordination always entails a temporal component, the Cumulative-cue hypothesis stays agnostic of how precise timing relationships can contribute to prominence. Such a measure of alignment can only exist with both modalities present together and does not address their parallel or independent behavior. Thus, the present study can offer a new perspective to the Cumulative-cue framework on the precise temporal alignment of prosody and gesture as a form of coordination and how this alignment can signal prominence.

### *2.6. Motivation and research questions*

As discussed in the previous sections, while there is a body of research on the prosodic marking of focus as well as on the general coordination of prosody and gestures in terms of the presence and distribution of cues, less is known about the multimodal expression of pragmatic functions by means of the temporal integration of multimodal cues. Especially the variation in temporal coordination in pragmatic prominence contexts requires more attention, for which information-structural contexts provide the ideal linguistic environment. The present study addresses two research questions, one concerning unimodal prosodic marking of focus and a second one on the multimodal temporal marking of focus.

Regarding the prosodic analysis, several studies have found prosody to vary as a function of focus in German (see Kügler & Calhoun, 2020 for an overview), showing that prosodic prominence increases in more prominent pragmatic

contexts. What is unknown for German, however, is whether the temporal alignment of falling pitch accents with the segmental string varies as a function of focus like in other languages (e.g., Kohler, 2005; Hanssen et al., 2008; Ladd & Morton, 1997). Therefore, we investigate the temporal prosodic marking of focus. We test this to assess whether hyperarticulation (Lindblom 1990) can be found in prosodic terms in this dataset, in line with previous studies (Hanssen et al., 2008). Concretely, we address the following research question: (Q1) *Do the slopes of falling accents differ as a function of focus in German?* We hypothesize that the slope of falling accents is steeper in focus than it is in background as a result of more precise (laryngeal) articulation, in line with previous findings for West Germanic languages.

Studies on the multimodal marking of focus have mostly concentrated on the presence and distribution of gestural cues (Im & Baumann, 2020; Esteve-Gibert et al., 2022) and found that higher pragmatic prominence is associated with a higher frequency of gestures. We investigate the variation of temporal alignment of prosodic and gestural landmarks as a function of focus. We extend the idea of variation in temporal alignment in focus marking to multimodal cues, taking a novel approach as temporal alignment of prosodic and gestural landmarks has not been studied regarding how it varies in focus marking contexts. This study thus contributes to gaining a deeper understanding of the prosody-gesture link as well as multimodal focus marking. Concretely, we address the following research question: (Q2) *Does the temporal alignment of gestural and prosodic landmarks vary as a function of focus?* We hypothesize that, similarly to purely prosodic focus marking, alignment of multimodal cues is tighter in focus than it is in background, as a result of a more precise articulation of both prosodic and gestural landmarks in contexts of higher pragmatic prominence.

## 3. General methods of the corpus study

To address these questions, we conducted a corpus analysis using the Bielefeld Speech and Gesture Alignment corpus (SaGA corpus, Lücking et al., 2010, 2013), investigating the relation between prosodic landmarks and syllables on the one hand (unimodal perspective), and the prosodic and gestural landmarks (multimodal perspective) on the other hand in a conversational setting. We consider prosody to be one modality of communication, in which we investigate the temporal behavior of prosodic landmarks from a unimodal perspective. With gestures representing another modality of communication, the alignment analysis investigates the coordinated behavior of the two modalities regarding their temporal alignment from a multimodal perspective.

### 3.1. The dataset: SaGA corpus and data preparation

The SaGA corpus is an audiovisual corpus containing 25 dialogues (280 min) of German spontaneous speech and was recorded in Bielefeld (Lücking et al., 2010, 2013). Two interlocutors perform a task of giving and following directions (spatial communication) in an unplanned, but task-directed, conversation. The direction description task serves the purpose of eliciting gestures, since speaking about a spatial topic has been shown to elicit more gestures than discussing abstract or verbal topics (Alibali, 2005). In the task scenario, one participant (describer) receives a tour through a virtual reality town using VR glasses prior to a recorded conversation. During the tour, the describer makes mental notes of five characteristic landmarks (sculpture, town hall, church square, chapel, fountain) as well as the route connecting them. The describer then explains the tour and landmarks to the second interlocutor (listener), who is supposed to find their own way through the VR-town afterwards. The conversation itself is the target of the corpus, not the successful task completion. An illustrative sketch of the task underlying the SaGA corpus data elicitation is provided in Fig. 4.

Every conversation in the SaGA corpus was recorded with three cameras from three different perspectives (full describer view, full listener view, total view of conversation). In addition to a separate audio file, the corpus contains several layers of annotation. Lexical annotations are provided on the sentence level in ELAN (Wittenburg et al., 2006; ELAN [Computer Software], 2024). Regarding gestures, multiple physical characteristics of the gestures are provided in the corpus, as well as gesture types and gesture phrases (i.e., a bigger gestural domain, including potential preparation or recovery movements), but gesture phases (gestures split into stroke and preparation, recovery, hold separately) are provided for iconic gestures only. Gesture types are largely annotated following McNeill (1992), distinguishing between 'iconic' and 'deictic' gestures as well as 'discourse' gestures and 'beat' gestures. As the latter two both refer to short, abstract and often rhythmic gestures which are not semantically connected to speech, we consider the latter two gesture types to be non-referential gestures following more recent gesture classification systems (e.g., M3D; Rohrer et al., 2025).

Using a spontaneous speech corpus for this study increases ecological validity. The speech and behavior of the participants are more natural and spontaneous than with sentences collected in a controlled experimental setting (for an example of a very controlled setting, see Roustan & Dohen, 2010). Therefore, a near-to-natural utilization of gestures and accentuation can be expected.

### 3.2. Participants and conversations

For the present study, of the 25 SaGA corpus dialogues, those with available audio, video and annotation files were

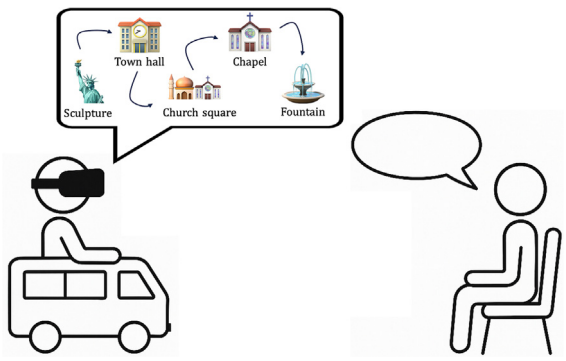


**Fig. 4.** Illustrative sketch of the SaGA corpus elicitation task. A describer explains the route of a bus tour through a VR-town environment with five characteristic landmarks to a listener, who then has to find their way through the town by themself.

used. After excluding seven dialogues due to missing data, we analyzed 18 dialogues (204 min) for their prosodic and gestural behavior in focus and background contexts. The speech and gestures of both conversation partners were analyzed, resulting in 36 participants in total. The average duration of a conversation was 11:33, with the shortest conversation being 5:18 and the longest 19:26. The participants were 16 female and 20 male speakers (although the majority of describers - with the major speaking part - were male; n = 14/18). The corpus does not provide any further personal or demographic information about the participants. However, judging from the videos, most participants seem to be in the age span of 25–35 years, and since the corpus was recorded in Bielefeld, North-Western German most likely was the predominant regional dialect.

### 3.3. Annotations and inter-rater reliability

To address our research questions, we added annotations on the focus dimension of information structure, prosody in terms of pitch accent types and gesture apex annotations.

The dimension of information structure corresponding to focus and background was annotated by the first author according to the LISA guidelines (Götze et al., 2007) relying on text annotations only, classifying focused constituents as information or contrastive focus and distinguishing these from background constituents. A constituent was labeled as information focus if it contained the most important or new sentence information. It was labeled as contrastive focus if it (explicitly) evoked contrast to another element. As an example, in (1), the sentence from the SaGA corpus contains background (B), information focus (IF) and contrastive focus (CF) constituents based on the sentence context:

(1) Context [. . .] der Allee folgen und, nee, am Rathaus war das rechts, genau rechts. [. . .]
*[. . .] follow the alley and, no, at the town hall it was to the right, yes to the right. [. . .]*
Target [Wenn du rechts abgebogen bist]$_{B}$ [eins, zwei Kurven]$_{IF}$, [also einmal]$_{B}$ [stark links]$_{CF}$ [und dann]$_{B}$ [sofort rechts]$_{CF}$.
*[When you have turned right]$_{B}$ [one, two curves]$_{IF}$, [so once]$_{B}$ [sharply left]$_{CF}$ [and then]$_{B}$ [immediately right]$_{CF}$.*

All remaining non-focus speech constituents were labeled as background constituents. Identifiable silences (speech breaks), disfluencies in speech and unattached filler particles were not included in the background constituents and thus not considered for analysis. The size of the constituents was variable. However, constituent length in the SaGA corpus was similar among the focus types: background constituents had a mean duration of 0.73 sec (median = 0.65 sec), information focus had a mean duration of 0.70 sec (median = 0.65 sec) and contrastive focus had a mean of 0.73 sec (median = 0.67 sec). Maximum duration differed more, as the maximum background length was 4.22 sec and the maximum information focus length was 3.60 sec, while the maximum contrastive focus length was 2.10 sec.

The SaGA corpus contains gesture annotation in ELAN (Wittenburg et al., 2006; ELAN [Computer Software], 2024), which was done based on the annotation system that Kopp and colleagues (2004) developed and which is in some aspects similar to the CoGEST (Trippel et al., 2004) and the FORM (Martell et al., 2002) annotation schemes. In the present study, we analyzed non-referential (comprising 'beat' and 'discourse' gestures from the corpus) and iconic gesture apexes, all of which were uni-directional. The corpus already contained annotations of iconic gesture strokes, but strokes of non-referential gestures were added for this study by the first author following the MultiModal MultiDimensional annotation system (M3D, Rohrer et al., 2025). Apexes of both iconic and non-referential gestures were also annotated for this study following M3D. An apex is characterized as the peak of a gestural stroke or, in other words, the kinematic goal of the movement. An apex can be identified by the sharpness of the picture frame in the video, since the apex is a point of minimum velocity, represented by little or no movement. All gesture annotations were performed in a video-only setting, thus without access to the audio.

For prosody, pitch accents were annotated in an audio-only setting by the first author following GToBI (Grice et al., 2005) in Praat (Boersma & Weenink, 2024). They were assigned to the corresponding accented syllable and aligned with the relevant f0-maximum or f0-minimum. For the prosodic analysis of falling pitch accents, specifically H+L* and H+!H* accents were considered (see section 4.1 below), as these have been found to be associated with focus and show variation in different focus types (in Dutch, Hanssen et al., 2008). For the multimodal analysis, pitch accent types were not distinguished as all types of pitch accents associate with gestures (e.g., Im & Baumann, 2020). However, annotating pitch accent types is relevant for deciding the positioning of the accents, as pitch accents were used as the starting point to identify f0-maxima and minima which in turn are the reference points for measuring and calculating temporal alignment.

To assess the reliability of annotations, a part of the database was annotated by a second annotator, and we ran Cohen's Kappa over the annotations. For focus, constituents were segmented by the main annotator with the opportunity for the second annotator to disagree and re-segment, and 6637 constituents (13.3%) were labelled by two annotators for focus condition. Agreement for focus was high (N = 6637, $\kappa$ = 0.783, z = 78.6, p < 0.001). For gestures, 421 apexes (17%) were annotated by two annotators for their placement and gesture type (iconic or non-referential) based on the existing stroke annotation. Agreement was very high (N = 421, $\kappa$ = 0.805, z = 18.6, p < 0.001). For prosody, pitch accents in one conversation were annotated by two annotators independently, and 327 pitch accents (4.2%) were matched on the same word. Accents were grouped according to their tonal target, H* (includes H*,!H* and L+H*), or L* (includes L* and L*+H), with falling accents assessed as a separate category due to the relevance for the analysis. There was substantial agreement for pitch accents (N = 327, $\kappa$ = 0.75, z = 14.7, p < 0.001). This indicates a good reliability across all annotation dimensions.

### 3.4. General statistics

The dataset contains 10,268 constituents, which are distributed as follows: 6171 background constituents and 4097 focus constituents, which comprise 3363 (82.1%) information focus constituents and 734 (17.9%) contrastive focus constituents. To account for the imbalance in the distribution of focus constituents, which naturally emerges from the spontaneous speech corpus, we used Bayesian modeling for inferential statistics. The dataset contains 7795 pitch accents, of which 5052 accents were produced on focus constituents and 2743 were produced on background constituents.

For this study, two analyses were performed, a prosodic analysis of the SLOPES of falling contours (see details in section 4.1) and a temporal alignment analysis in terms of the DISTANCE between apexes and pitch accents (see details in section 5.1). For both analyses, the predictor variable was FOCUS, distinguishing between information focus, contrastive focus and background constituents. Statistical analysis was performed with R, version 4.4.2 (R Core Team, 2024). We used the following R packages: *tidyverse,* version 2.0.0 for data wrangling (Wickham et al., 2019), *ggplot2*, version 3.5.1 (Wickham, 2016) and *patchwork*, version 1.3.0 (Pedersen, 2024) for plotting and *brms*, version 2.22.0 (Bürkner, 2017), *bayesplot*, version 1.11.1.9000 (Gabry & Mahr, 2024) and *emmeans*, version 1.10.6 (Lenth, 2024) for Bayesian modeling. The detailed models used for both analyses are presented in the data processing sections, 4.1 and 5.1. The R script used for the analysis can be found on OSF: https://osf.io/5tzk2.

## 4. The prosodic analysis

The prosodic analysis addresses the research question (Q1), *whether the slopes of falling accents differ as a function of focus in German* from a unimodal perspective.

### 4.1. Data processing and statistics

The full dataset contains 390 falling pitch accents (5% of all accents). For each constituent that was marked by a falling accent, we annotated the accented syllable with the falling accent (which contained the local f0-minimum) and its preceding syllable (which usually contained the local f0-maximum) which formed the two target syllables for each item. We considered the two types of falling accents H+L* and H+!H* together, as H+!H* rarely occurred (two in background, four in information focus and one in contrastive focus). We also only included falling contours as part of pitch accents, excluding falling contours consisting of an H* and boundary tone. We then extracted the f0 at ten points per syllable from the sound file using the Praat script ProsodyPro (Xu, 2013) to retrieve time-normalized f0-contours. From these points, we identified the f0-maxima and f0-minima of the contour. The f0-slope was calculated as the f0-difference between the maximum and minimum divided by the time difference between the time of the f0-maximum and that of the f0-minimum.

Of the original 390 data points, 16 (4.1%) had to be excluded due to pauses between the first and second target syllable, and 29 (7.4%) additional data points were excluded as outliers since their slope was outside the range of 25 and 750 Hz/sec. The upper slope limit of 750 Hz/sec emerged as this limit is two standard deviations (2 × 252 Hz/sec) away from the mean slope (244 Hz/sec) of the condition with the lowest mean and standard deviation (background), meaning that it excludes the most peripheral 5% of the data in the upper limit. Below 25 Hz/sec, the slope was considered too flat to present a noticeable difference. This resulted in a total of 45 items (11.5%) being excluded from the analysis, and 345 items being included.

We ran Bayesian lognormal models (*brms*, Bürkner, 2017) to test the steepness of the SLOPE of falling accent contours as a function of FOCUS. We tested background against information focus and contrastive focus. Since lognormal models exclusively run over positive values, we added a constant of 0.0001 Hz/sec to all absolute slope values to compute the model. None of the models indicated any convergence issues (Rhat values = 1.00 and no divergent transitions). As priors for the probability distribution, we used weakly informative priors with a logarithmic normal distribution and default priors for intercept and standard deviation. We ran the models with 8 Markov Chain Monte Carlo (MCMC) chains with 8000 iterations each (2000 warmups), which resulted in 48,000 post-warmup samples. We assessed the model fit with posterior predictive checks, and we computed estimated marginal means on the log-scale and exponential values to assess the difference between the focus distributions. Effects were considered robust if the 95%-credible intervals (henceforth 95%-CrI) did not include 0. The models for the prosodic analysis do not include any random effects (for participants and items) because the models with random effects did not converge due to the limited size of the dataset (Bulk ESS was too low and Rhat value = 1.15). We analyzed raw slope values (Hz/sec) and those normalized to semitones/sec, but since the analyses did not show differences in the results, we report the raw slopes for reasons of readability.

### 4.2. Results of the prosodic analysis

The following section presents the results of the prosodic analysis of the SLOPES of falling contours across the background and two focus conditions. In total, 345 falling accents comprise 197 (57.1%) information focus accents, 54 (15.7%) contrastive focus accents (251 focus accents in total) and 94 (27.2%) background accents and their SLOPES are compared. Fig. 5 displays the time-normalized mean f0-contours of falls based on the f0-points per syllable (dashed lines) in the three conditions across the syllable containing the minimum of the fall and its preceding syllable. The full line represents the smoothed contour.

The following can be observed: The contours in background and information focus are alike in many aspects, only diverging slightly towards the end of the second syllable, and information focus overall has a slightly higher f0. In contrast, the falling contour of the accents in contrastive focus differs clearly, which is most visible in the height of the pitch peak in the first syllable, before f0 drops down to the same height as the other conditions in the second syllable. The high peaks (maximum f0) of contrastive focus are on average 25 Hz higher than those of information focus and background. Concretely, the mean maximum f0 of contrastive focus is at 211 Hz, while it is at 187 Hz in

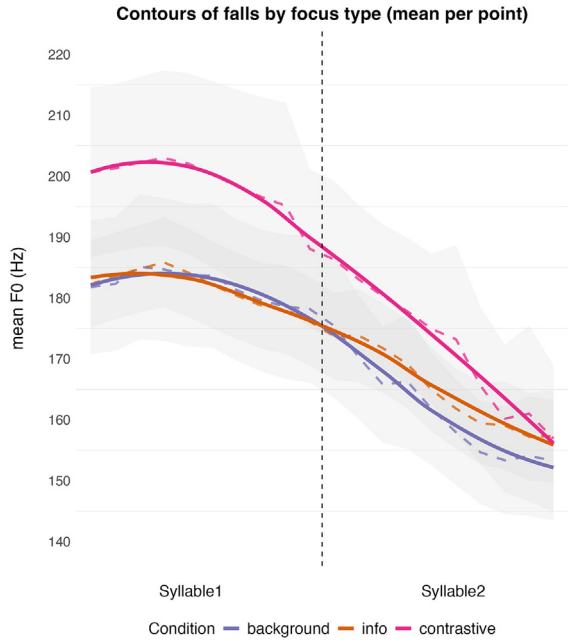


**Fig. 5.** Mean f0-contours of falling accents in background, information focus and contrastive focus on the target syllables. The dashed lines represent the contours based on ten f0-measurements per syllable; the solid lines are smoothed.

background and 196 Hz in information focus. The mean f0-minimum of all conditions is at a similar height, being 142 Hz in background, 146 Hz in information focus and 151 Hz in contrastive focus. Table 1 provides an overview of the relevant mean Hz measures.

A closer look at the slopes of the three conditions shows that falling accents on contrastive constituents have the steepest fall (247 Hz/sec) in comparison to falling accents in background (198 Hz/sec) and information focus (182 Hz/sec), which has the flattest slope. The greater fall and steeper slope in contrastive focus arises from a higher f0-maximum and similar f0-minimum. Violin plots in Fig. 6 show the distribution and means of the slopes of falling accents for each focus condition, illustrating the steeper slopes in contrastive focus and the similarity of distributions in background and information focus.

The Bayesian lognormal models indicate an effect of focus on the slope of falling contours between background and contrastive focus (background: $\beta = 5.01$, 95%-CrI [4.87, 5.16]; contrastive: $\beta = 0.27$, 95%-CrI [0.03, 0.51]) and between information focus and contrastive focus (information: $\beta = 4.97$, 95%-CrI [4.87, 5.07]; contrastive: $\beta = 0.31$, 95%-CrI [0.09, 0.52]) with the credible intervals not including 0. There is no difference in the slopes between background and information focus (information: $\beta = 4.97$, 95%-CrI [4.87, 5.07]; background: $\beta = 0.04$, 95-%CrI [-0.14, 0.22]) with the credible intervals including 0. This indicates an effect of FOCUS on the SLOPE of falling accents as the slopes of falling accents are steeper in contrastive focus than they are in background and information focus.

Transforming the log-values of the Bayesian models back to the original slope values (in Hz/sec), the model predicts the following slopes for each of the conditions: 150 Hz/sec (95%-CrI [129, 172]) for background, 144 Hz/sec (95%-CrI [131, 160]) for information focus and 197 Hz/sec (95%-CrI [161, 236]) for contrastive focus. The model prediction underlines the steeper slopes of falling contours in contrastive focus, which are steeper by 46 HZ/sec (95%-CrI [92.3, 4.55]) compared to background and by 52 Hz/sec (95%-CrI [94.7, 13.41]) compared to information focus. Fig. 7 (left) illustrates the posterior distributions for each focus condition, where a small or no overlap of the distributions indicates a reliable difference, while a considerable overlap shows the similarity of the posterior distributions, which indicates that there is no difference between the conditions. Fig. 7 (right) shows the differences between focus types as predicted by the model, with the 95%-credible intervals excluding 0 being an indication of a reliable difference predicted by the model.

### *4.3. Interim summary*

The results of the prosodic analysis reveal that the slopes of falling contours vary in different focus contexts in German spontaneous speech. The slopes in contrastive focus are systematically steeper than they are in background and information focus contexts. This difference in slopes is contingent on a higher f0-maximum and a fall to a similar f0-minimum within the scope of two syllables in contrastive focus compared to the other conditions. The steeper fall can be interpreted as more precise articulation of the f0-contour in the contrastive focus context, as a bigger f0-fall is produced during the same time frame. However, due to the difference between contrastive focus and information focus but not between information focus and background, this does not appear to be a direct and general marker of focus in this dataset, but specifically a marker of contrast.

The effects found in this analysis are reliable but not very strong, which may be due to the data being based on spontaneous speech — which naturally varies more than controlled lab speech (e.g., De Ruiter, 2015), and can less well be controlled for balance of items per condition, which is in particular visible in our contrastive focus condition having only few falls. In addition, the corpus contained a limited number of falling accents that were relevant for this analysis. The fact that an effect of (contrastive) focus on the slopes of falling contours can be observed nevertheless supports the idea of adapting articulation as a function of focus.

**Table 1**
Overview of the relevant mean Hz measures in the prosodic analysis, per condition (background, information focus, contrastive focus).

| Condition | Number of items | f0-Max mean | f0-Min mean | f0-Fall mean | f0-Slope mean |
|---|---|---|---|---|---|
| Background | 94 | 187 Hz | 142 Hz | 45 Hz | 198 Hz/sec |
| Information focus | 197 | 194 Hz | 146 Hz | 47 Hz | 182 Hz/sec |
| Contrastive focus | 54 | 211 Hz | 151 Hz | 60 Hz | 247 Hz/sec |

## 5. The prosody-gesture alignment analysis

The analysis addresses the research question (Q2), *whether the temporal alignment of gestural and prosodic landmarks varies as a function of focus* from a multimodal perspective.

### 5.1. Data processing and statistics

For every apex in the corpus, the nearest pitch accent was identified without further reference to any semantic connection, as the connection of interest was on a temporal level. The only restriction was that both had to be on the same focus or background constituent (see Türk & Calhoun, 2024 for a similar approach to investigating prosody-gesture interaction in the same information-structural domain). In cases where multiple gesture apexes are present in the same domain, the apex closest to the accent was chosen. Thus, each item consists of a unique apex–accent pair, for which the distance between the apex and f0-extremum was measured in milliseconds. A distance of 0 ms means that apex and extremum of the accent occur exactly at the same time. Fig. 8 illustrates an example of the prosodic and gestural annotations performed for this purpose, merged in ELAN, and shows how the annotations of prosody and gesture are temporally related.

To investigate the average distance between apexes and accents, we focused on the absolute distance between the two landmarks instead of considering raw distances, which take into account which of the two landmarks precedes the other. The sample was then split by FOCUS to compare the distances of the apex-accent alignment between background, information focus and contrastive focus. Pitch accent types were not considered, as all types of pitch accents may be associated with a gesture. The annotation of pitch accent types was still relevant, as the type of pitch accent influences the temporal positioning of the extremum and therefore accent annotation; H+L*, for instance, has L* as a temporal reference, while L+H* has H* as the reference.

In total, 2472 relevant gesture apexes (iconic or non-referential) were found in the dataset. Of these, 1801 were considered for analysis, since they formed a pair with an accent that was situated on the same focus or background constituent. 671 apexes (27.1%) were excluded from the analysis in total: 309 gesture apexes (12.5%) were not produced on speech constituents, but during silences or disfluencies, and 362 apexes (14.6%) were produced during speech, but not on a constituent bearing a pitch accent. Regarding gesture rates, the general occurrence of gestures relative to pitch accents is comparable to previous studies (Im & Baumann, 2020; Kügler & Gregori, 2023; Rohrer et al., 2023), with this dataset containing a ratio of approximately one gesture to three pitch accents.

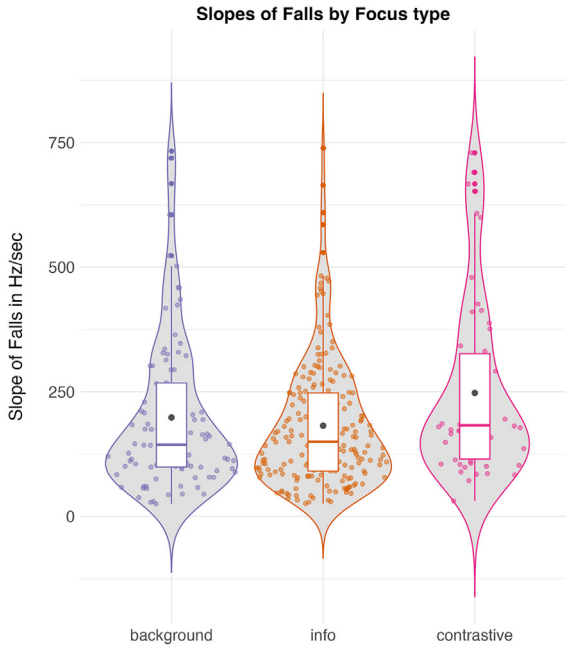


**Fig. 6.** Violin plots of the slopes of falling contours in background, information focus and contrastive focus. The plot shows the distribution of falls as well as the mean and median in Hz/sec.

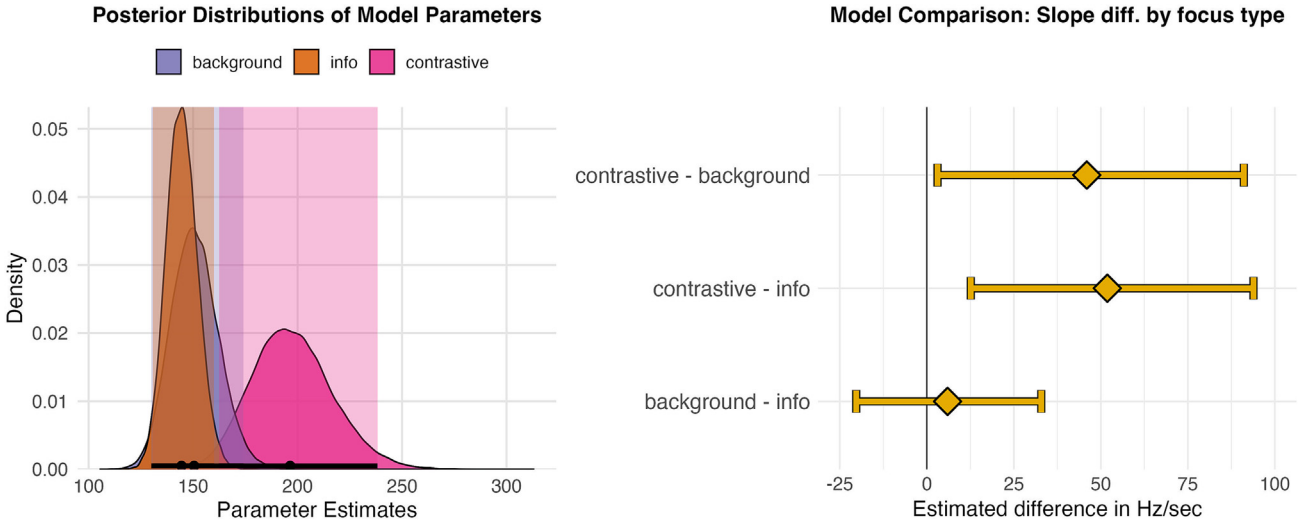


**Fig. 7.** Posterior distributions for each focus condition (left) and estimated differences between focus conditions (right) as predicted by the lognormal models of slopes of falling contours.

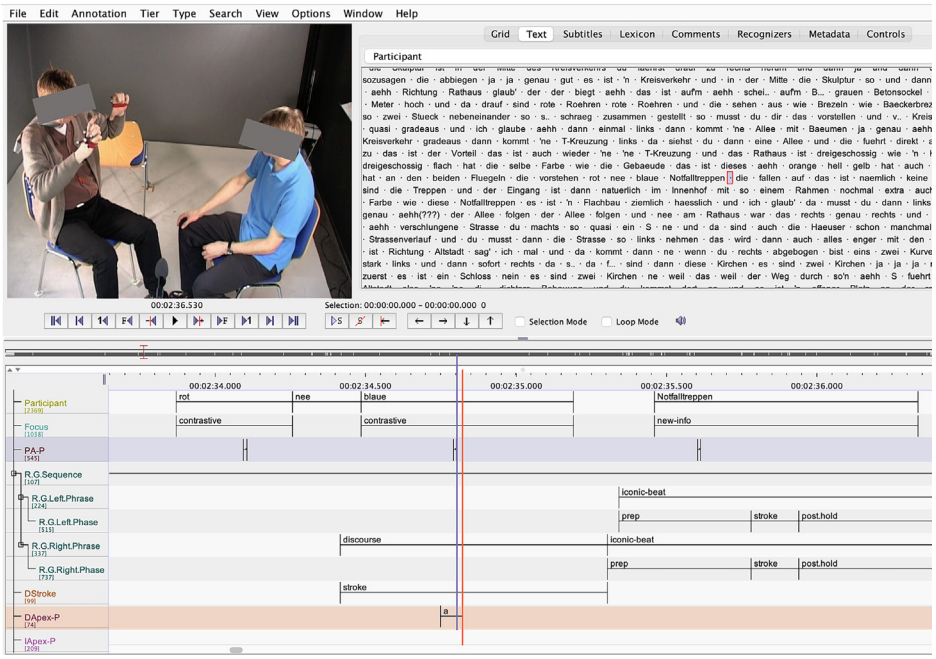


**Fig. 8.** Illustration of prosodic and gestural annotations in ELAN of an apex-accent pair with very close temporal alignment. The orange tier and orange vertical bar indicate the position of the gesture apex; the purple tier and purple vertical bar indicate the position of the associated pitch accent. Since ELAN only allows for annotation in intervals, the corresponding right boundaries of said intervals are aligned with the apex and accent positions and taken as points of reference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We ran Bayesian lognormal models (*brms*, Bürkner, 2017) to test the temporal alignment in terms of the absolute DISTANCE between gesture apexes and pitch accents as a function of FOCUS. We ran a model testing background against information focus and contrastive focus. Since lognormal models exclusively run over positive values and the current dataset includes one apex-accent pair with a distance of exactly 0 ms, we added a constant of 0.0001 ms to all absolute distance values to compute the model. We added random intercepts for PARTICIPANTS and ITEMS to the models. None of the models indicated any convergence issues (Rhat values = 1.00 and no divergent transitions). As priors for the probability distribution, we used weakly informative priors with a logarithmic normal distribution and default priors for intercept and standard deviation. We ran the models with 8 Markov Chain Monte Carlo (MCMC) chains with 8000 iterations each (2000 warmups), which resulted in 48,000 post-warmup samples. We assessed the model fit with posterior predictive checks, and we computed estimated marginal means on the log-scale to assess the difference between the focus distributions. Effects were considered to be robust if the 95%-credible intervals did not include 0.

### 5.2. Results of the prosody-gesture alignment analysis

The following section presents the results of the multimodal temporal alignment analysis of the DISTANCE between apexes and pitch accents. 1801 apexes were included in the analysis – 675 on background constituents (37.5%), 954 on information focus (53%) and 172 on contrastive focus (9.5%). In general, apexes and accents occurred temporally associated with each other, as displayed in the absolute distances in the apex-accent pairs in Fig. 9. A distance of 0 ms indicates that the apex and accent occur at exactly the same time. The mean absolute distance in apex-accent pairs is 227 ms (SD = 243 ms). The mean distance between apexes and accents is displayed as the dashed yellow line. A particular clustering of pairs at the lower end of the scale is visible.

In terms of temporal alignment split by FOCUS, the results show that apexes and accents align more closely in focus than in background contexts, as displayed in Fig. 10. The distribution of distances in apex-accent pairs is clearly tighter in both focus types than it is in background and the proportion of tightly aligning pairs is higher in focus than in background. The mean absolute distance is 328 ms (SD = 326 ms) in background, opposed to 169 ms (SD = 148 ms) in information focus and 153 ms (SD = 125 ms) in contrastive focus. Contrastive focus has a slightly lower distance than information focus. In terms of maximal distance in an apex-accent pair, the picture is similar, as in background, the maximum distance is 2399 ms, in information focus it is 980 ms and in contrastive focus it is 604 ms.

We ran a Bayesian lognormal model over the DISTANCE in apex-accent pairs as a function of FOCUS. The model shows an effect of FOCUS on the absolute distance between apex and accent: both information and contrastive focus have a reliably lower distance between apexes and accents than background (background: $\beta = 5.27$, 95%-CrI [5.17, 5.37]; information: $\beta = -0.62$, 95%-CrI [−0.74, −0.50]; contrastive: $\beta = -0.75$, 95%-CrI [−0.95, −0.55]), with the credible intervals being far away from and not including 0. The models do not indicate reliable differences in distances between the focus types (information: $\beta = 4.65$, 95%-CrI [4.56, 4.74]; contrastive: $\beta = -0.13$, 95%-CrI [−0.33, 0.07]) with the credible intervals including 0. See

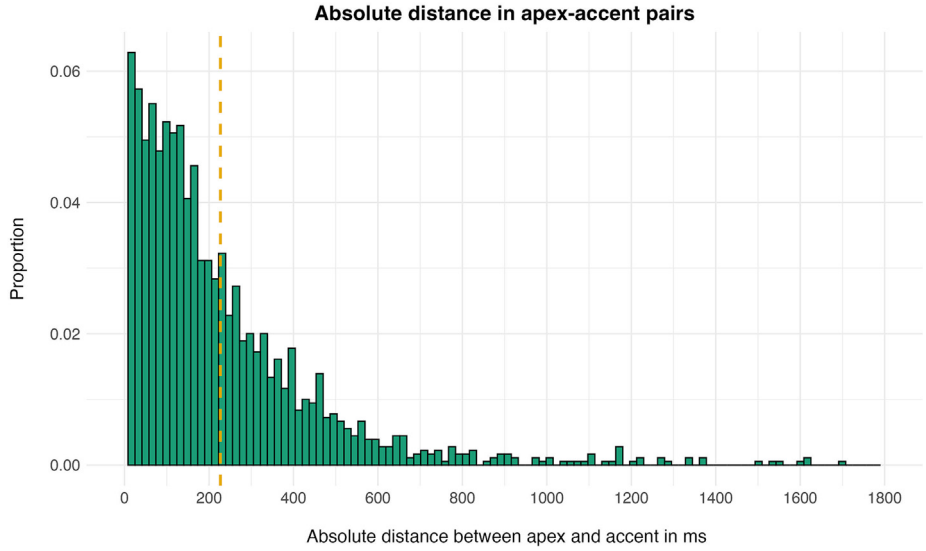


**Fig. 9.** Proportions of absolute distances in apex-accent pairs. The dashed yellow line indicates the mean distance between apexes and accents. The x-axis displays values up to 1800 ms distance. A few apex-accent pairs with distances up to 2400 ms were found, which are not displayed in this plot for reasons of clarity and comparability of the figures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

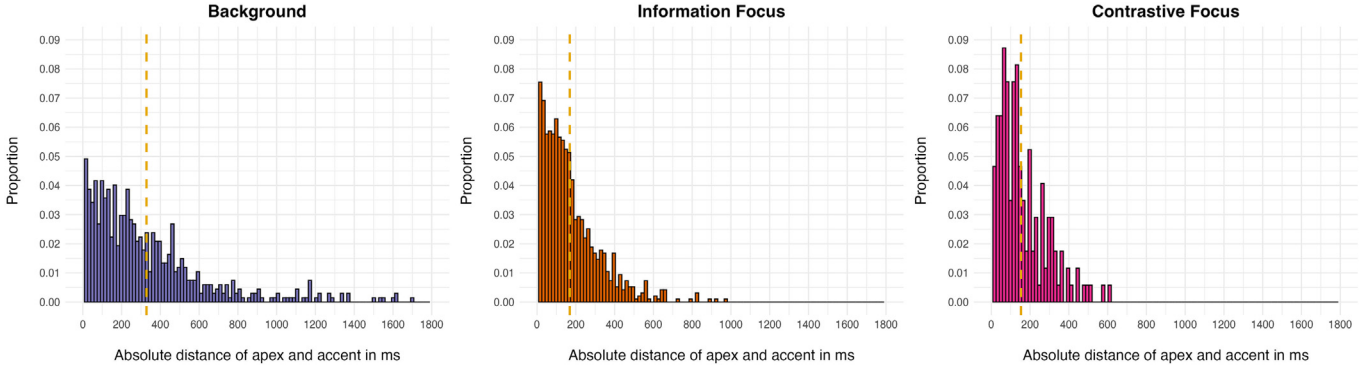


**Fig. 10.** Proportions of absolute distances in apex-accent pairs split by Focus (background, information focus, contrastive focus). The dashed yellow line indicates the mean distance between apexes and accents for each distribution. The x-axis displays values up to 1800 ms of distance for all conditions. In background, a few apex-accent pairs with distances up to 2400 ms were found, which are not displayed in this plot for reasons of clarity and comparability of the figures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

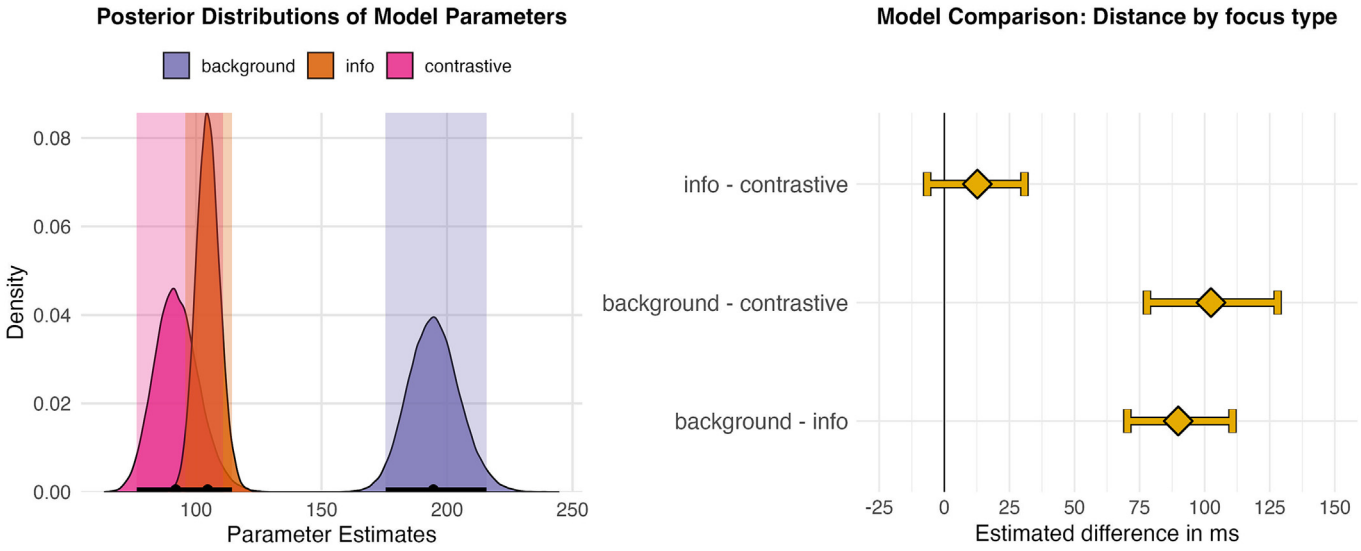


**Fig. 11.** Posterior distributions for focus conditions (left) and estimated differences between focus types and background (right) as predicted by the lognormal models of absolute distances in apex-accent pairs.

Fig. 11 (left) for the posterior distributions. This indicates that the apex–accent distance is lower on focus constituents than it is on background constituents. Since there is no overlap between the background distribution with the focus types, this effect can be considered to be strong and reliable.

Transforming the log-values of the model back to original measures of distances in milliseconds indicates that the model predicts a mean distance of 195 ms (95%-CrI [174.8, 215]) for background, a mean of 105 ms (95%-CrI [95.6, 114]) for information focus and a mean of 92 ms (95%-CrI [75.5, 110]) for contrastive focus. The model predictions underline the reliable difference in alignment between background and focus constituents with a distribution difference of 90 ms (95%-CrI [70.3, 110.7]) between background and information focus and 102 ms (95%-CrI [77.8, 128.0]) between background and contrastive focus, as illustrated in Fig. 11 (right).

### 5.3. Interim summary

The results of the multimodal analysis indicate that temporal alignment of prosodic and gestural landmarks varies as a function of focus. Generally, pitch accents and gesture apexes are temporally associated with each other (with a mean distance of 227 ms), which indicates that apexes and accents tend to surround the same word. The distance between pitch accents and gesture apexes is smaller in focus than it is in background contexts, and thus their alignment is more precise. Temporal alignment does not differ reliably across focus types, which shows that the alignment difference boils down to a distinction between focus and background. As the mean constituent duration does not differ between focus types in the whole dataset, the differences in temporal alignment are associated with the pragmatic focus context rather than a general constituent length effect.

## 6. Discussion

This corpus study investigated the role of temporal alignment from both a unimodal and a multimodal perspective in focus marking in German spontaneous conversations. From the unimodal perspective, the prosodic analysis of falling F0 slopes revealed steeper falls in contrastive focus compared to other contexts. From the multimodal perspective, gesture-prosody alignment was closer in focus constituents than in background constituents. In the following sections, the results will be discussed in detail (Sections 6.1 and 6.2 respectively) and evaluated in terms of Hypo- and Hyperarticulation theory (Section 6.3).

### 6.1. Discussion of prosodic findings

The prosodic analysis was concerned with the variation in f0 of falling pitch accents on words in focus and background and concretely addressed the research question (Q1), *whether the slopes of falling accents differ as a function of focus in German.* The results in background, information focus and contrastive focus revealed that the slopes are steeper in contrastive focus than they are in background and information focus. Steeper slopes in contrastive focus are driven by a higher f0-maximum on the pre-stressed syllable and a similar f0-minimum in the stressed syllable compared to the other two contexts. These findings are in line with previous findings on the prosodic marking of contrast (e.g., Baumann et al., 2007; Braun, 2006; Féry & Kügler, 2008), and with results obtained from a perceptual perspective (Kügler & Gollrad, 2015).

Furthermore, the evidence that contrastive focus is marked by steeper falling f0-slopes, mainly caused by differences in f0-maxima, indicate a variation in effort in different prominence contexts (e.g., Baumann et al., 2007; Féry & Kügler, 2008). This finding is in line with H&H-theory (Lindblom, 1990) and the Effort Code (Gussenhoven, 2002, 2004), as higher pragmatic prominence causes speakers to use more articulatory modulation and precision to highlight important information (e.g., Repp, 2016; Cruschina, 2021). Note, however, that in this study's data, this effect is only visible in the most prominent contexts (contrastive focus), while the slopes of information focus falling contours rather resemble background contours.

This resemblance of slopes in background and information focus is not in line with previous findings that background and focused information are distinguished from each other in prosody. It has been found numerous times that background and given information is more likely to be deaccented or reduced in prominence, while focus is often highlighted (e.g., Baumann & Hadelich, 2003; Baumann et al., 2006, 2007). Supporting the expected difference between background and focus constituents, a relative positioning effect is known for West Germanic languages: Background or given constituents often appear in pre- or post-focal position, and accents in this position have been claimed to be weaker. In particular, pre-focal accents are described as ornamental (non-functional or non-obligatory, Büring, 2007; but see Roessig, 2023) or rhythmic (Calhoun, 2010), post-focal constituents are often reduced or deaccented (Ladd, 2008; Kügler & Féry, 2017 among many others). The similarity between background and information focus in our study may likely be related to the task, which required memorizing a route. Hence, many repetitions yielding highlighting of given information are contained in the corpus because the describer presumably wants to make sure that the listener remembers the route through the town. This memorizing effect may override a general less prominent marking of background, resulting in less distinguishable intonation of background and information focus. Contrastive focus on the other hand occurred in contexts where explicit comparisons were drawn or differences between features were explained (e.g., *'[the left church]$_{CF}$ had a round roof, [the right church]$_{CF}$ had two towers'*), leading to a more distinct intonation with regard to f0-contours, which is in line with the literature (Kügler & Calhoun, 2020 for an overview).

Another possible explanation for the strong difference in the slopes between information focus and contrastive focus could be related to the type of falling accent primarily used in the respective categories. Grice et al. (2009) report on differences in tonal alignment between H+!H* and H+L* falling accents, which are clearly observable, as the Hz-difference between the local f0-maxima and f0-minima is smaller for the H+!H* accent than for the H+L* accent. For our data, this is however an unlikely explanation, as H+!H* rarely occurred in the data and no systematic differences between focus types could be determined based on this accent type distinction. Therefore, we believe the steepness of slopes to be a context-dependent effort rather than a distinct tonal specification.

Overall, the prosodic analysis showed that speakers produce a steeper slope on falling contours in contexts in which the conveyed information is more relevant for the comprehension of the listener (cf. Lindblom, 1990). As this steeper slope is produced in a comparable time frame to slopes on background information (namely the stressed syllable and its preceding syllable), we interpret this observation in terms of H&H-theory as hyperarticulation by higher accuracy in timing pitch accents in focused contexts. While similar results of an increase in precision and effort have been shown in articulatory cases (e.g., that prominence increases clearness of sound articulation, De Jong, 1995), the results of this study indicate such an increase in laryngeal terms, which we interpret as close alignment on the contrastively focused constituent since there is more f0-movement in the same time frame than in other conditions. This result and interpretation are in line with findings on the slopes of falling contours by Hanssen and colleagues (2008) for Dutch, who found more precisely aligned intonational contours in narrow (information and contrastive) focus than in broad focus, i.e., in a bigger focus domain. Note that our study used a different scale for focus types. While Hanssen et al. (2008) used a 'broad focus' condition as a baseline to compare more prominent focus realizations, our study used background constituents for comparison. In our study, which investigated precisely the time window of the stressed syllable and the preceding one, the steeper slopes are found only in the marking of contrastive focus compared to background but not for the marking of information focus. The study on Dutch found a more precise alignment of f0-extrema with the stressed syllable in narrow focus than in broad focus. While the scope of focus was not the subject of investigation in the present study, the steeper slopes in the very narrow time frame of two syllables makes it likely that this finding could be found in German data as well. With the pattern of steeper slopes of falling contours in contrastive focus contexts found in this dataset, we interpret this as an instantiation of prosodic hyperarticulation in German spontaneous speech. Concretely, regarding the first research question (Q1), this study concludes that slopes of falling contours are steeper in focus than in background, but only in a contrastive context. We interpret this as hyperarticulation because the contrast that the speaker aims to convey to the listener leads to a higher articulatory effort.

### 6.2. Discussion of prosody-gesture alignment findings

The multimodal analysis was concerned with the temporal alignment of prosody and gesture and concretely addressed the research question (Q2), *whether the temporal alignment of gestural and prosodic landmarks varies as a function of focus.* Specifically, the analysis of the temporal distance between gesture apexes and associated pitch accents in different focus conditions revealed that prosody and gesture are aligned more closely in focus than they are in background, with no reliable difference in alignment between information and contrastive focus. This finding provides evidence for a tighter alignment of prosody and gesture in pragmatically more prominent contexts.

This dataset shows general temporal alignment of gesture apexes and pitch accents (see Fig. 9), which generally supports McNeill's (1992) claim of the temporal coordination of prosody and gesture. This finding is in line with other empirical evidence on the temporal alignment of prosody (accents and syllables) and gestures (apexes and strokes) (e.g., Shattuck-Hufnagel et al., 2007; Loehr, 2012; Esteve-Gibert & Prieto, 2013; Shattuck-Hufnagel & Ren, 2018). The results of our study do not make any reference to the order of appearance of gestural and prosodic cues. Whereas McNeill's phonological synchrony rule assumes that gestures precede or occur at the same time as stressed syllables, our data do not indicate which cue comes first due to the absolute distance measure. As we do not integrate any semantic restrictions for association of the cues, a flexible ordering is possible in this study. However, addressing the order of prosodic and gestural cues is beyond the scope of this study.

Regarding the effect of focus on apex-accent alignment, the results showed that the alignment was narrower (meaning a smaller distance between the landmarks) on focus than on background constituents. This difference is very clear and can be seen in the tighter distribution of apex-accent differences in both focus types and the broader distribution in background (see Fig. 10), as well as in the mean distance, which was about half as big in focus than it was in background. This is true for both information focus and contrastive focus, whereas the temporal alignment did not differ between the focus types. We interpret this smaller distance between apexes and accents as precise alignment and therefore more precise articulation in focus than in background. In focus, gestures are temporally associated with prosodic landmarks, demonstrating a more precise multimodal articulation in contexts which are more relevant to the listener. These results are in line with previous studies finding a more precise alignment of gestural with prosodic cues in focus (Lehečková et al., 2022; Kügler & Gregori, 2023; Türk & Calhoun, 2024), indicating temporal alignment to be a marker of prominence. The present study differs from Lehečková et al. (2022), who found alignment differences between information and contrastive focus in Czech. This difference may be attributed to the imbalanced dataset in the present study, which had relatively fewer contrastive focus apex-accent pairs than information focus ones. It could also emerge from different analysis strategies: while this study investigated the absolute distances between prosodic and gestural landmarks, the Czech study took order of the cues into account and analyzed the onset of the gesture, finding that contrastive focus strokes are produced earlier. Since the onset of the gesture was analyzed in the Czech study, an earlier gestural stroke does not necessarily lead to a closer alignment with prosody. The general trend of a closer alignment in focus compared to background is however confirmed across all studies.

The results of the multimodal analysis support more general findings that prosody and gesture are involved in focus marking (e.g., Baills & Baumann, 2025; Coego et al., 2025; Ebert et al., 2011; Esteve-Gibert et al., 2022; Gregori et al., 2024), all of which indicate that the two modalities use increased cues with higher contrastivity. This parallel increase leads to an increase in the impression of multimodal prominence, which is addressed by the Cumulative-cue hypothesis (cf. Ambrazaitis & House, 2022, 2023). The Cumulative-cue hypothesis assumes a parallel increase of prosodic and gestural cues in different dimensions to increase the multimodal

prominence impression. The present study did not investigate the potential parallel patterns of prominence in prosody and gesture, but it took a measure of distance, which can only exist when multiple modalities are used and combined. Thus, while these data do not directly reflect the cumulation aspects in the Cumulative-cue hypothesis, and the hypothesis stays agnostic about temporal aspects, the present results show coordination of prosody and gesture, as well as a change of this coordination for the purpose of prominence, which extends on the Cumulative-cue hypothesis in the temporal domain.

Parallel to the unimodal perspective in the prosodic analysis, we interpret the results of the multimodal analysis as an instantiation of hyperarticulation: The pragmatic context of focused constituents, where speakers identify a need to highlight information results in a closer temporal association of gesture apexes and pitch accents, which we interpret as hyperarticulated speech. More cognitive effort is put in by the speaker to achieve a more precise and simultaneous production of speech and gesture, which is done with the intention of emphasizing information for the hearer in multimodal terms. The observations on the temporal alignment of prosody and gesture answer the second research question (Q2), concluding that temporal alignment is more precise in focus than it is in background.

### 6.3. The findings in terms of H&H-theory

This study considers hyperarticulation a communicative phenomenon concerning emphasis and precision of articulation that goes beyond the acoustic modality of speech, where H&H-theory has been established. Hyperarticulation concerns the segmental level (Lindblom, 1990), but also the prosodic level, as studies on intonational contours (Hanssen et al., 2008) and this study's prosodic analyses of the slopes of f0-falls show. This study's multimodal analysis, which evaluated the timing of prosodic and visual landmarks in relation to each other, provided evidence for hyperarticulation, as posited by H&H-theory, being used in a multimodal setting. The data on the distance between gestural and prosodic landmarks in focus are an instantiation of multimodal hyperarticulation (in background multimodal hypoarticulation). As illustrated in Fig. 12 (which expands on Fig. 1), the general concept of hyperarticulation predicts a more precise articulation, which results in a tighter distribution of a measured variable, e.g., formant frequencies of vowels. In more specific uni- or multimodal hyperarticulation, this precision of articulation can be expressed in different forms, as for instance in the steepness of slopes of falling contours or the temporal alignment of gestural and prosodic landmarks. Note, that other measures are expected to display multimodal hyperarticulation or unimodal hyperarticulation on the prosodic (e.g., intensity, duration, spectral features) or the gestural level (e.g., salience, acceleration, number of articulators) and the analyses of this study work as examples for the respective modes of hyperarticulation.

Approaching hyperarticulation as a multimodal phenomenon highlights the simultaneity and similarity of the pragmatic functions that prosody, gesture and their interaction have in communication. First steps in exploring multimodal hyperarticulation have also been taken in sign language research (e.g., Mertz, Pagel, Perniss, et al., 2024; Mertz, Pagel, Turco, et al., 2024), but more research is needed. The marking of information structure, and focus specifically, is a linguistic function well suited for investigating hyperarticulation, as these contexts provide the environment for pragmatic prominence by calling for highlighting, making hyperarticulation more likely. At the same time, note that the notion of focus is orthogonal to hyperarticulation, as focus is highly dependent on the context and also allows for non-enhanced or reduced productions (i.e., hypoarticulation) in cases where competing information is not temporally close in discourse or is no likely alternative. The fact that there are apex-accent pairs in the data that are relatively far away from each other in both background and focus additionally highlights the orthogonal concept of hyperarticulation and prominence. From this, we can deduce that focus is a domain that is likely to lead to hyperarticulation,

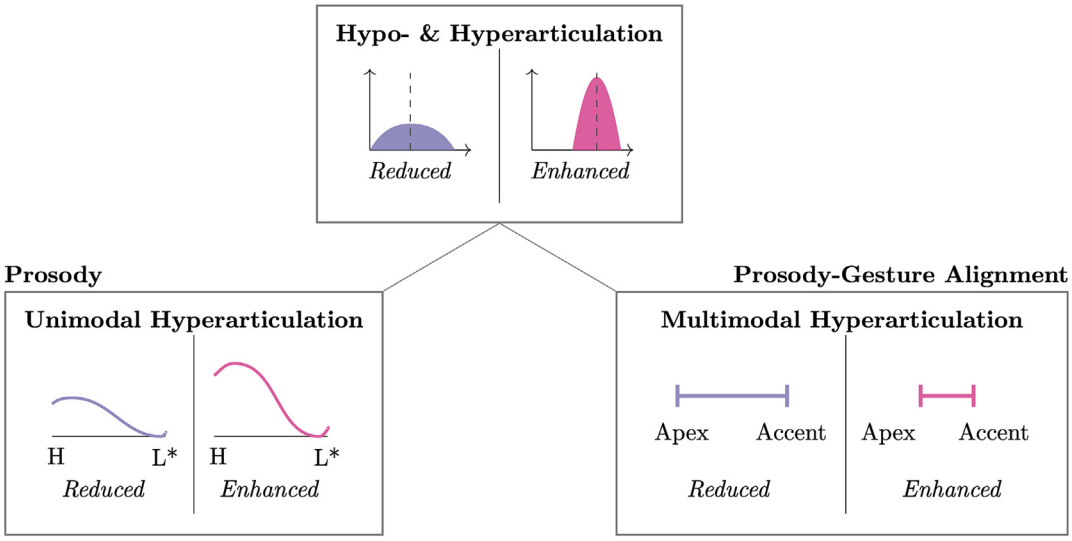


**Fig. 12.** Schematic illustration of hypo- and hyperarticulation in general, of unimodal hypo- and hyperarticulation on the prosodic level and of multimodal hypo- and hyperarticulation on the temporal alignment level.

but does not mandatorily have to, rendering this relation probabilistic. Identifying temporal alignment of prosody and gesture as multimodal hyperarticulation in the expression of information structure contributes to research on the prosody-gesture-pragmatics interface. This can be expanded further by exploring other instances of multimodal hyperarticulation, not only related to temporal alignment, but involving both phonetic and kinematic features that can be connected in the acoustic and visual channels (e.g., amplitude, acceleration or velocity of sound and gesture). The use of these features is subject to the Cumulative-cue hypothesis, which states that the coordination of acoustic and visual cues takes place in parallel, leading to multimodal prominence.

Importantly, the deviation of results between the prosodic and multimodal analyses, namely that prosodic hyperarticulation distinguishes information and contrastive focus while multimodal hyperarticulation distinguishes background and focus, highlights the importance of the interplay of prosody and gesture in communication. As this study investigated a phenomenon that is not existent within the channel of (spoken) language only, this highlights the fundamental role of gestures in communication. Gestures can contribute to the distinction of pragmatic contexts (here focus types) that are not distinguished by other domains like syntax or prosody (cf. Debreslioska & Gullberg, 2020; Rohrer, 2022). The interaction of prosodic and visual cues marks focus, in addition to the prosodic channel. Changes in alignment are more pronounced in the multimodal data, where the reliable differences between focus conditions are bigger (i.e., distances between apex and accent are almost 50% smaller in focus, while f0-slope is about 20% steeper in contrastive focus). As discussed above, the missing distinction between background and information focus in the prosodic channel may be a task effect which is connected to memory enhancement. Memory enhancement refers to a general function of prosody, i.e., highlighting information (Gussenhoven, 2004; Ladd, 2008), which may explain the highlighted background constituents despite many of them being given in discourse. Parallel operation of prosody and gesture could lead to the expectation that gestural marking would also not be reduced in background due to the memory demand. However, this is not what the data show. An explanation for this could be that while repeated items may be reinforced acoustically, this may not be true for corresponding gestural cues, with gestures being either less frequent or produced with less precision, restricting the marking of repetition to one channel only. This flexible behavior of the modalities is in line with the modality-neutral prosodic framework hypothesis (Prieto et al., 2024), stating that modalities need not enhance in parallel but may be expressed individually to reach a particular communicative goal (cf. Trujillo & Holler, 2023; Żygis & Fuchs, 2023).

The deviating results between the analyses further show that both unimodal and multimodal hyperarticulation need to be considered to achieve a full distinction between the three investigated focus conditions, as both marking strategies only make a binary distinction: the prosodic analysis discriminates contrastive contexts from non-contrastive contexts, while the multimodal analysis discriminates background and focus. Neither strategy distinguishes all three contexts, which is only achieved by considering both strategies in combination. This highlights the benefits and importance of considering unimodal and multimodal cues in the marking of pragmatic functions.

### 6.4. Limitations and future work

While this study gave insight into unimodal prosodic alignment and multimodal alignment of prosody and gesture, and its effects in background and focus, some (widely methodological) limitations arise. First and foremost, the present study is based on corpus data, which was not explicitly designed for this study. Using an existing corpus has many advantages (e.g., data partially prepared for analysis), but it allows less control of the data. The underlying task of the corpus (route description and memory task) may have influenced the gestures that were used, as it may lead to many spatial gestures. In addition, the task was demanding with substantial repetitive conversation involved, which may influence information structure produced in discourse. In particular, the background condition mentioned material that would be considered given information but that was more highlighted than usual to keep the interlocutor's attention. An analysis of segmental reduction on the repeated tokens could shed light on variation in the strength of the background condition, adding an independent acoustic perspective to the prosodic analysis. Finally, another potential influencing factor on the prosodic analysis that could not be controlled for were prosodic phrasing, sentence position or accent status. These factors are known to affect the prominence of an accent and the timing of the f0-extrema (e.g., Esteve-Gibert & Prieto, 2013; Surányi, 2024), especially since these factors are at interplay with information structure and should be controlled in future studies where applicable.

Further, this study zoomed in on very specific prosodic and multimodal features that, while being relevant to prominence and to the precision of articulation, cannot represent the entire domain. Hyperarticulation concerns a broad range of cues, which need to be addressed in future work. On the prosodic side, not only should other pitch accent types and f0-contours beyond falling accents be considered, but other phonetic cues like duration, intensity or tonal center of gravity are known to vary in linguistic contexts (e.g., Kochanski et al., 2005; Kügler, 2008) and may thus be relevant in hyperarticulated speech. On the gestural side, a variety of kinematic features should be examined for their variation along the hypo- and hyperarticulation continuum, including gesture amplitude, velocity or acceleration. From a multimodal perspective, other forms of coordination between prosody and gesture (e.g., magnitude) could be hyperarticulated. Even within temporal alignment, alternative pairing criteria for accents and gestures could be considered, referring to different prosodic and gestural anchors or different decisions regarding pairings across focus constituent boundaries. While this study imposed the restriction for the prosodic and gestural cue to occur in the same focus domain and otherwise focusing on their temporal connection, this opens the possibility for a close apex-accent pair to not be paired when they cross a constituent boundary, which in turn could lead to a higher distance-pairing to emerge, resembling a spill-over effect. While this decision was made in accordance with the research question, future studies could consider different criteria, which may affect or further refine the results.

## 7. Conclusion

This study investigated the prosodic and multimodal marking of focus by means of temporal features in a corpus of German multimodal conversational data, showing that the steepness of slopes of falling accents as well as the temporal alignment of gesture apexes and pitch accents are more precise in focus than they are in background contexts. Following hypo- and hyperarticulation theory (Lindblom, 1990), we interpret both results as instantiations of hyperarticulation. Prosodically, we confirmed hyperarticulation in falling contours as a function of contrastive focus marking. The concept of multimodal hyperarticulation is supported by the observation that prosody and gesture work together to distinguish focus constituents from background constituents. Only by combining prosodic and multimodal hyperarticulation could a full distinction between background, information focus and contrastive focus be achieved. Multimodal hyperarticulation is a concept that should be further explored, emphasizing various gestural features, to assess the similarity of prosodic and multimodal processes in communication.

### CRediT authorship contribution statement

**Alina Gregori:** Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Frank Kügler:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

### Data availability

Part of the Bielefeld SaGA corpus is publicly available in the 'Bayrisches Archiv für Sprachsignale' (BAS), however this does not include all dialogues used in this study. As the administration and management of the corpus take place at Bielefeld University (Stefan Kopp and colleagues), the videos and annotations for the present study cannot be made publicly available. We are happy to provide the annotations for the present study upon request. The R script for data processing, inter-rater reliabilities and analysis is available in this OSF: https://osf.io/5tzk2.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

We thank Andy Lücking for providing the SaGA corpus and Nathalie Klenner for assistance with data annotation. Thanks to Aleksandra Ćwiek for her advice on Bayesian model fitting and thanks to Paul Koenig for his help compiling Figs. 1 and 12. We thank three anonymous reviewers for their very valuable and encouraging feedback and Caroline Féry for commenting on an earlier version of this paper. Thanks to Kirsten Brock for English proofreading.

We would like to acknowledge the financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – [project ID 502013510, KU 2323/5-1 & PR 1780/1-1]; SPP 2392 Visual Communication (ViCom), Frankfurt am Main, Germany – and Goethe University Frankfurt.

### References

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition & Computation, 5*(4), 307–331. https://doi.org/10.1207/s15427633scc0504_2.

Ambrazaitis, G., & House, D. (2022). Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters. *Laboratory Phonology, 24*(1). https://doi.org/10.16995/labphon.6430.

Ambrazaitis, G., & House, D. (2023). The multimodal nature of prominence: Some directions for the study of the relation between gestures and pitch accents. In O. Niebuhr (Ed.), *Proceedings of the 13th International Conference of Nordic Prosody* (pp. 262–273). Sciendo. https://doi.org/10.2478/9788366675728-024.

Atterer, M., & Ladd, D. R. (2004). On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. *Journal of Phonetics, 32*(2), 177–197. https://doi.org/10.1016/S0095-4470(03)00039-1.

Baills, F., & Baumann, S. (2025). Prosody and head gestures as markers of information status in French as a native and foreign language. *Language and Cognition, 17*, e42. https://doi.org/10.1017/langcog.2025.11.

Baumann, S., Becker, J., Grice, M., & Mücke, D. (2007). Tonal and articulatory marking of focus in German. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1029–1032.

Baumann, S., Grice, M., & Steindamm, S. (2006). Prosodic marking of focus domains-categorical or gradient. *Proceedings of Speech Prosody, 2006*, 301–304.

Baumann, S., & Hadelich, K. (2003). Accent type and givenness: An experiment with auditory and visual priming. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*, 1811–1814.

Baumann, S., & Röhr, C. T. (2015). The perceptual prominence of pitch accent types in German. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*. In 18th International Congress of Phonetic Sciences (ICPhS), Glasgow, Scotland, 0384.

Beckman, M. E. (Ed.). (2010). *Stress and non-stress accent*. Foris.

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer [Computer program]* (Version Version 6.4.04) [Computer software]. http://www.praat.org/.

Braun, B. (2006). Phonetics and phonology of thematic contrast in German. *Language and Speech, 49*(4), 451–493. https://doi.org/10.1177/00238309060490040201.

Büring, D. (2007). Intonation, semantics and information structure. In G. Ramchand, & C. Reiss (Eds.), The Oxford Handbook of Linguistic Interfaces (pp. 445–474). Oxford, OUP.

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01.

Calhoun, S. (2010). How does informativeness affect prosodic prominence? *Language and Cognitive Processes, 25*(7–9), 1099–1140. https://doi.org/10.1080/01690965.2010.491682.

Chafe, W. (1976). Giveness, contrastiveness, definiteness, subjects and topics. In C.N. Li (ed) Subject and Topic, New York: Academic Press, 27–55.

Coego, S., Esteve-Gibert, N., & Prieto, P. (2025). Preschoolers mark focus types through multimodal prominence: further evidence for the precursor role of gestures. *Languages, 10*(5), 92. https://doi.org/10.3390/languages10050092.

Cruschina, S. (2021). The greater the contrast, the greater the potential: On the effects of focus in syntax. *Glossa: A Journal of General Linguistics, 6*(1), 3. https://doi.org/10.5334/gjgl.1100.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America, 97*(1), 491–504. https://doi.org/10.1121/1.412275.

De Ruiter, L. E. (2015). Information status marking in spontaneous vs. read speech in story telling tasks – Evidence from intonation analysis using GToBI. *Journal of Phonetics, 48*, 29–44. https://doi.org/10.1016/j.wocn.2014.10.008.

Debreslioska, S., & Gullberg, M. (2020). What's new? Gestures accompany inferable rather than brand-new referents in discourse. *Frontiers in Psychology, 11*, 1935. https://doi.org/10.3389/fpsyg.2020.01935.

Ebert, C., Evert, S., & Wilmes, K. (2011). Focus marking via gestures. *Proceedings of Sinn & Bedeutung, 15*(15), 193–208.

*ELAN [Computer software]* (Version 6.7). (2024). [Computer software]. The Language Archive. https://archive.mpi.nl/tla/elan.

Elsner, A. (2000). *Erkennung und Beschreibung des prosodischen Fokus*. (PhD Thesis). Rheinische Friedrich-Wilhelms-Universität Bonn. https://hdl.handle.net/20.500.11811/1663.

Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., & D'Imperio, M. (2022). Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Developmental Science, 25*(1). https://doi.org/10.1111/desc.13154 e13154.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research, 56*(3), 850–864. https://doi.org/10.1044/1092-4388(2012/12-0049).

Féry, C. (1993). *German intonational patterns*. Tübingen: M. Niemeyer.

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics, 36*(4), 680–703. https://doi.org/10.1016/j.wocn.2008.05.001.

Fung, H. S. H., & Mok, P. P. K. (2018). Temporal coordination between focus prosody and pointing gestures in Cantonese. *Journal of Phonetics, 71*, 113–125. https://doi.org/10.1016/j.wocn.2018.07.006.

Gabry, J., & Mahr, T. (2024). *bayesplot: Plotting for Bayesian Models* (Version 1.11.1.9000) [R package]. https://mc-stan.org/bayesplot/.
Gerwing, J., & Bavelas, J. (2005). Linguistic influences on gesture's form. *Gesture, 4*(2), 157–195. https://doi.org/10.1075/gest.4.2.04ger.
Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S., & Stoel, R. (2007). Information structure. *Interdisciplinary Studies on Information Structure: ISIS, 7*, 147–187.
Gregori, A., Amici, F., Brilmayer, I., Ćwiek, A., Fritzsche, L., Fuchs, S., Henlein, A., Herbort, O., Kügler, F., Lemanski, J., Liebal, K., Lücking, A., Mehler, A., Nguyen, K. T., Pouw, W., Prieto, P., Rohrer, P. L., Sánchez-Ramón, P. G., Schulte-Rüther, M., ... Von Eiff, C. I. (2023). A roadmap for technological innovation in multimodal communication research. In V. G. Duffy & V. Duffy (Eds.). *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management* (Vol. 14029, pp. 402–438). Switzerland: Springer Nature.
Gregori, A., Sánchez-Ramón, P. G., Prieto, P., & Kügler, F. (2024). Prosodic and gestural marking of focus types in Catalan and German. *Proceedings of Speech Prosody, 2024*, 891–895 10.21437/SpeechProsody.2024-180.
Grice, M., Baumann, S., & Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology. In S.-A. Jun (Ed.), Prosodic Typology. The Phonology of Intonation and Phrasing (pp. 55–83). Oxford: OUP.
Grice, M., Baumann, S., & Jagdfeld, N. (2009). Tonal association and derived nuclear accents—The case of downstepping contours in German. *Lingua, 119*(6), 881–905. https://doi.org/10.1016/j.lingua.2007.11.013.
Grice, M., & Kügler, F. (2021). Prosodic prominence – A cross-linguistic perspective. *Language and Speech, 64*(2), 253–260. https://doi.org/10.1177/00238309211015768.
Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. *Proceedings of Speech Prosody, 2002*, 47–57 10.21437/SpeechProsody.2002-7.
Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.
Hanssen, J., Peters, J., & Gussenhoven, C. (2008). Prosodic effects of focus in Dutch declaratives. *Proceedings of Speech Prosody, 2008*, 609–612. https://doi.org/10.21437/SpeechProsody.2008-138.
Im, S., & Baumann, S. (2020). Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America, 5*(1), 685–697. https://doi.org/10.3765/plsa.v5i1.4755.
Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication, 25*(1980), 207–227.
Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
Kohler, K. (2005). Timing and Communicative Functions of Pitch Contours. *Phonetica, 62*(2–4), 88–105. https://doi.org/10.1159/000090091.
Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*(2), 1038–1054. https://doi.org/10.1121/1.1923349.
Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. *Proceedings of the 6th International Conference on Multimodal Interfaces*, 97–104. https://doi.org/10.1145/1027933.1027952.
Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica, 55*(3–4), 243–276. https://doi.org/10.1556/ALing.55.2008.3-4.2.
Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology, 8*(1). https://doi.org/10.5334/labphon.75.
Kügler, F. (2008). The role of duration as a phonetic correlate of focus. *Proceedings of Speech Prosody, 2008*, 591–594. https://doi.org/10.21437/SpeechProsody.2008-134.
Kügler, F., & Calhoun, S. (2020). Prosodic Encoding of Information Structure: A typological perspective. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 453–467). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198832232.013.30.
Kügler, F., & Féry, C. (2017). Postfocal downstep in German. *Language and Speech, 60*(2), 260–288. https://doi.org/10.1177/0023830916647204.
Kügler, F., & Gollrad, A. (2015). Production and perception of contrast: The case of the rise-fall contour in German. *Frontiers in Psychology, 6*, 1254. https://doi.org/10.3389/fpsyg.2015.01254.
Kügler, F., & Gregori, A. (2023). Iconic gestures in focus – Synchronization of prosody and gestures in prominence. *Proceedings of the 20th International Congress of Phonetic Sciences*, 4125–4129.
Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684.
Ladd, D. R. (1980). *The structure of intonational meaning: Evidence from English*. Indiana University Press.
Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.
Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics, 25*(3), 313–342. https://doi.org/10.1006/jpho.1997.0046.
Ladewig, S. (2011). Syntactic and semantic integration of gestures into speech: structural, cognitive, and conseptual asects. (PhD thesis). Universität Frankfurt/Oder.
Lambrecht, K. (2000). Information structure and sentence form: Topic, focus, and the mental representations of discourse referents (Repr., digitally print. [der Ausg.] 1998). Cambridge Univ. Press.
Lehečková, E., Jehlička, J., & Králová Zíková, M. (2022). Multimodal marking of information structure: Gesture-prosody alignment across languages. *Linguistica Pragensia, 32*(1), 19–38 10.14712/18059635.2022.1.2.
Lenth, R. (2024). *_emmeans: Estimated Marginal Means, aka Least-Squares Means_* (Version 1.10.6) [R package]. https://CRAN.R-project.org/package=emmeans.
Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218.
Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Netherlands: Springer. https://doi.org/10.1007/978-94-009-2037-8_16.
Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture, 7*(2), 179–214. https://doi.org/10.1075/gest.7.2.04loe.
Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1). https://doi.org/10.1515/lp-2012-0006.
Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces, 7*(1–2), 5–18. https://doi.org/10.1007/s12193-012-0106-8.
Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In M. Kipp, J.-P. Martin, P. Paggio, & D. Heylen (Eds.), *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality* (pp. 92–98).
Martell, C., Osborn, C., Friedman, J., & Howard, P. (2002). The FORM Gesture Annotation System. In *Proceedings of the Multimodal Resources and Multimodal Systems Evaluation Workshop* (pp. 15–22).
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
McNeill, D. (2006). Gesture: A psycholinguistic approach. *The Encyclopedia of Language and Linguistics, 1*. 85–66.
Mertz, J., Pagel, L., Perniss, P., Turco, G., & Mücke, D. (2024a). Coarticulation in sign language: A kinematic study on French Sign Language (LSF) using Electromagnetic Articulography (EMA). *ISSP 2024–13th International Seminar on Speech Production*, 51–54.
Mertz, J., Pagel, L., Turco, G., & Mücke, D. (2024b). Gradiency and categoriality in the prosodic modulation of French Sign Language: A kinematic approach using Electromagnetic Articulography. *Proceedings of Speech Prosody, 2024*, 851–855. https://doi.org/10.21437/SpeechProsody.2024-172.
Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America, 96*(1), 40–55. https://doi.org/10.1121/1.410492.
Niebuhr, O., & Kohler, K. J. (2004). Perception and cognitive processing of tonal alignment in German. *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 155–158.
McNeill, D., & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill, (Ed.), *Language and gesture* (Vol. 1987, pp. 141–161). Cambridge: CUP.
Pedersen, T. (2024). *_patchwork: The Composer of Plots_* (Version 1.3.0) [R package]. https://CRAN.R-project.org/package=patchwork.
Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology, 9*, 1109. https://doi.org/10.3389/fpsyg.2018.01109.
Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 28*(1), 96–103. https://doi.org/10.1044/jshr.2801.96.
Pouw, W., De Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2021). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences, 1491*(1), 89–105. https://doi.org/10.1111/nyas.14532.
Pouw, W., & Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science, 43*(3). https://doi.org/10.1111/cogs.12721 e12721.
Prieto, P., Esteve-Gibert, N., & Shattuck-Hufnagel, S. (2024). Towards a novel conceptualization of prosody that accounts for spoken and visual signals: The modality-neutral prosodic framework hypothesis. *Gesture* (1-2), 119–159. https://doi.org/10.1075/gest.25012.pri.
R Core Team. (2024). *R: A Language and Environment for Statistical Computing. [Programming language]* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.
Repp, S. (2016). Contrast: Dissecting an Elusive Information-structural Notion and its Role in Grammar. In C. Féry & S. Ishihara (Eds.), *The Oxford Handbook of Information Structure* (1st ed., pp. 270–289). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642670.013.006.
Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics, 5*, 1–69. https://doi.org/10.3765/sp.5.6.
Roessig, S. (2023). Prosody of pre-focal background depends on following focus. *Proceedings of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic*, 1220–1224.
Rohrer, P. L. (2022). *A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system* (PhD Thesis). Nantes Université; Universitat Pompeu Fabra (Barcelone, Espagne).

Rohrer, P. L., Delais-Roussarie, E., & Prieto, P. (2023). Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses. *Lingua, 293*, 1–17. https://doi.org/10.1016/j.lingua.2023.103583.

Rohrer, P. L., Hong, Y., & Bosker, H. R. (2024). Gestures time to vowel onset and change the acoustics of the word in Mandarin. *Proceedings of Speech Prosody, 2024*, 866–870. https://doi.org/10.21437/SpeechProsody.2024-175.

Rohrer, P. L., Tütüncübasi, U., Florit-Pons, J., Vilà-Giménez, I., Esteve-Gibert, N., Ren-Mitchell, A., Shattuck-Hufnagel, S., & Prieto, P. (2025). Multidimensional labeling of gesture in communication: The M3D proposal. *Corpus Pragmatics*. https://doi.org/10.1007/s41701-025-00197-2.

Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. *Proceedings of Speech Prosody, 2010*(100110), 1–4.

Schlenker, P. (2019). Gestural semantics: Replicating the typology of linguistic inferences with pro- and post-speech gestures. *Natural Language & Linguistic Theory, 37*(2), 735–784. https://doi.org/10.1007/s11049-018-9414-3.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *The Journal of the Acoustical Society of America, 85*(1), 295–312. https://doi.org/10.1121/1.397737.

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology, 9*, 1514. https://doi.org/10.3389/fpsyg.2018.01514.

Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A method for studying the time alignment of gestures and prosody in American English: Hits' and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanić, E. Keller, & M. Marinaro (Eds.). *Fundamentals of verbal and nonverbal communication and the biometric issue* (Vol. 18, pp. 34–44). Amsterdam: IOS Press.

Sityaev, D., & House, J. (2003). Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In *Proceedings of the 15th ICPhS* (pp. 1819–1822).

Surányi, B. (2024). On the Role of Prosody in German Subpart of Focus Fronting. In A. Alexiadou, D. Georgi, F. Heck, G. Müller, & F. Schäfer, (Eds.), Gisbert Fanselow's Contributions to Syntactic Theory, Linguistische Arbeitsberichte 96, 271-286.

Trippel, T., Gibbon, D., Thies, A., Milde, J.-T., Looks, K., Hell, B., & Gut, U. (2004). CoGesT: A formal transcription system for conversational gesture. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2215–2218.

Trujillo, J. P., & Holler, J. (2023). Interactionally embedded gestalt principles of multimodal human communication. *Perspectives on Psychological Science, 18*(5), 1136–1159. https://doi.org/10.1177/17456916221141422.

Türk, O. (2020). *Gesture, prosody and information structure synchronisation in Turkish*. (PhD thesis).Victoria University of Wellington.

Türk, O., & Calhoun, S. (2024). Phrasal synchronization of gesture with prosody and information structure. *Language and Speech, 67*(3), 702–743. https://doi.org/10.1177/00238309231185308.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686 10.21105/joss.01686.

Wickham, H. (with Sievert, C.). (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer international publishing.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*(2), 159–197. https://doi.org/10.1016/j.wocn.2004.11.001.

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. *The Journal of the Acoustical Society of America, 115*(5_Supplement), 2397. https://doi.org/10.1121/1.4780717.

Zimmermann, M. (2008). Contrastive focus and emphasis. *Acta Linguistica Hungarica, 55*(3–4), 347–360. https://doi.org/10.1556/ALing.55.2008.3-4.9.

Żygis, M., & Fuchs, S. (2023). Communicative constraints affect oro-facial gestures and acoustics: Whispered vs normal speech. *The Journal of the Acoustical Society of America, 153*(1), 613–626. https://doi.org/10.1121/10.0015251.