ICONIC GESTURES IN FOCUS – SYNCHRONIZATION OF PROSODY AND GESTURES IN PROMINENCE

Frank Kügler & Alina Gregori

Institute for Linguistics, Goethe University Frankfurt, Germany
kuegler@em.uni-frankfurt.de, gregori@lingua.uni-frankfurt.de

ABSTRACT

This study investigates the impact of focus on the alignment of iconic gestures with prosody in spontaneous German speech. Generally, co-speech gestures synchronize with prosodic prominence [1]. From the SaGA Corpus [2], gesture types, iconic gesture apexes, and pitch accent types were extracted in focus and non-focus contexts. Results show that focused constituents were always marked with a pitch accent, and a minority were additionally accompanied by an iconic gesture. In 35% of unfocused constituents, no pitch accent was realized. Yet, the majority of unaccented constituents on non-focus constituents were accompanied by an iconic gesture. Regarding temporal synchronization, focus synchronizes the gestures more closely to the pitch accent than in non-focused constituents. This result points to the fact that a pragmatic function of highlighting seems to be added to iconic gestures to their otherwise ascribed dimension of expressing a semantic relation to speech concepts.

Keywords: Iconic co-speech gestures, prosody, focus, gesture-speech synchronization.

1. INTRODUCTION

This study investigates the impact of focus on the alignment of iconic co-speech gestures with prosody in spontaneous German speech. Co-speech gestures are assumed to accompany speech in a systematic way [e.g. 1, 3]. Related to prosody, a phonological synchrony rule between prosodic prominence and co-speech gestures has been claimed [1]: In general, it is assumed that a gesture coincides with a pitch accent, more specifically, the stroke of a gesture ends with the pitch accent at the latest. Basically, four different gesture types have been classified as iconic, deictic, metaphoric, and beat [1], the latter one being referred to as non-referential whereas the other three types are referential [4]. Since these types have different functions, the question arises whether all gesture types behave similar concerning the synchrony between prosodic prominence and gestures given (i) that different gesture types are assumed to be panned at distinct planning stages, e.g. [5].

For this study, we are concerned with iconic gestures, which show a semantic connection to speech in that they visually express the meaning of the co-occurring speech [1, 6]. See Figure 1 for an example of an iconic gesture where the speaker signals the shape of a pillar using his thumb and remaining fingers forming a cylinder. In what way the semantic function of iconic gestures interacts with prosodic prominence evoked by focus is the concern of the present study. To answer this, we conducted a corpus study on spontaneous German speech [2].

Figure 1: Example of an iconic gesture from the SaGA corpus [2]. The speaker utters the word *Säule* 'pillar', while imitating the shape of a pillar with his hands.

2. BACKGROUND

2.1. Information structure

We base our study on the information structure category *focus*, which according to [7, p. 247] "indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions." The non-focus part of an utterance is the labelled as *background* [8]. Languages use different linguistic means such as phonology, syntax, morphology or a combination thereof to express focus [9]. Stress-based languages such as Germanic languages use prosodic means, i.e. pitch accentuation to mark focus [10]. Hence, accentuation achieves the goal of making a focus more prominent than background information. Different types of focus are distinguished: Information focus is best illustrated in a question-answer pair where the answer identifies one of the alternatives being asked for. A contrastive focus usually includes a focus alternative, which was proposed in the immediately preceding context.

2.2. Prosody

In prosody, a distinction between prosodic domains according to the prosodic hierarchy [11, 12] and prosodic categories such as pitch accents, phrase and

boundary tones [13] is made, which we label using the GToBI system [14]. Pitch accents are usually the head of their prosodic domain [15] and have a highlighting function [13]. In German intonation, a number of different pitch accent types are assumed [14], and the pitch accent types bear different inherent prosodic prominence. According to [16], the pitch accents are organized on a prominence scale with rising or high pitch accents being more prominent than low ones.

2.3. Co-speech Gestures

Co-speech gestures are "visible bodily actions" accompanying speech [3]. Here, we concentrate on hand gestures. A widely used classification system of such gestures distinguishes four main types: iconic (Figure 1), metaphoric, deictic, and beat gestures [1]. A recent proposal groups the first three as referential gestures as opposed to non-referential ones, being assigned values in multiple dimensions [17]). We are interested in iconic gestures as they visually support and mirror the expressed semantic content of speech [1, 6 for an overview]. They thus allow for the transmission of additional or redundant information to the speech they accompany [6].

A gesture usually consists of multiple hierarchically ordered components [18]. Generally, the stroke of a gesture is assumed to align with syllables [1, 3] or pitch accents [17, 19–21]. Within the stroke, the apex of a gesture represents the gestural peak, which can be conceived of as the temporal point of no velocity, or the change of direction in the gesture's movement [4, 17]. Gestural strokes are integrated within a gesture phase, which in turn is integrated into a gesture phrase.

2.4. Prosody–Gesture–Link

According to [1], gestures and speech are two modalities of the same framework. For the integration of the two, he claimed three synchrony rules (pragmatic, semantic and phonological). For the latter one, he states that "the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech" [1, p. 26]. Later empirical work on the temporal synchronization of gestures and speech has shown that gestures and accents tend to occur near each other [19, 21]. The synchronization was observed for the level of the stroke and pitch accent [1], but also larger constituents such as gesture phrases and phonological phrases [21]. Recently, more empirical evidence has shown that information structure affects the synchronization [17, 22]. A referent carrying *new* information facilitates the occurrence of gestures.

2.5. Research question and hypothesis

The proposal of synchrony rules of gesture-speech integration [1] covers all gestures types. Focusing on iconic gestures, we are interested in their synchronization patterns with respect to prominence, i.e. focus. Given their function of mirroring the expressed semantic content of speech, the question arises if iconic gestures appear in focus and in particular, whether iconic gestures align closely with pitch accents in prominent position (focus).

Given the assumption that gestures can operate in multiple dimensions [23], we hypothesize that iconic gestures mark prominence in addition to their semantic contribution to speech and thus show sensitivity to pragmatic prominence, contemplating that iconic gestures occur more frequently and align more precisely with prosody in prominent positions.

3. CORPUS STUDY

3.1. Information on the corpus

The Bielefeld **S**peech **a**nd **G**esture **A**lignment (SaGA) corpus [2] is an audio-visual corpus containing German spontaneous speech conversations. The setting is based on a virtual reality (VR) town where participants were taken on a bus ride. The task was to inform an interlocutor about the route and certain landmarks (direction-giving task). In total, the data used from the corpus consists of 204 minutes of speech distributed over 18 dialogues. The corpus contained gesture type annotation according to [2].

3.2. Annotation and measurements

The corpus was further annotated for gesture apexes according to the M3D guidelines [4] using ELAN, for pitch accent types according to GToBI [14] using Praat [24], and for two levels of information structure (information status given, accessible, new; focus-background) [8]. Different types of focus (narrow information, contrastive focus) were collapsed for the analysis since prosodically, there were no significant differences in the prosodic realisation between focus types [e.g., 25].

From the corpus, pitch accent types (n = 4394), gesture occurrences (apex of non-referential and iconic gestures, n = 2402) and information structure levels (focus, n = 2773; background, n = 2251) were extracted. The present analysis focuses on iconic gestures (n = 1627). For the synchrony measure, the distance between the pitch accent and the iconic gesture's apex was calculated. Thus, each apex-accent pair was listed on a histogram according to their distance (in groups of one frame of 33ms length) The midpoint 0 ms means that apex and accent occur

at the same time. Positive values were assigned to the pairs in which the apex occurs before the accent, negative values when the accent occurs before the apex. Distances of more than 3000 ms were collapsed in a category 3000+.

4. RESULTS

4.1. Distribution of Pitch accents and Gestures

The distribution of pitch accents and iconic gestures split by focus is shown in Figure 2a. All focused referents (right bar) carry a pitch accent. Approximately 25% of the focused referents are additionally accompanied by an iconic gesture (red). Note that additionally about 10% are accompanied by non-referential gestures [26]. In the non-focus condition (background, left bar), approximately 63% of referents carry a pitch accent. From those, about 10% are additionally accompanied by iconic gestures (red). Approximately 30% of referents in non-focus carry no pitch accent but an iconic gesture (blue). For completeness, further 25% of non-focus constituents are accompanied by non-referential gestures (25% of these with an accent, 75% with no pitch accent) [26].

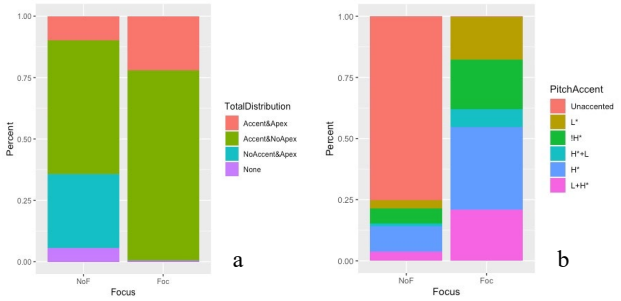

Figure 2: Distribution of pitch accents and iconic gestures split by focus (a); distribution of pitch accent types split by focus on iconic gestures (b).

4.2. Iconic gestures: Focus and Pitch accents

Zooming into the accent distribution on iconic gestures split by focus in Figure 2b (red and blue parts in Figure 2a): An iconic gesture in focus (right bar) occurs with all different pitch accent types of German. Note that we do not differentiate between focus types, and that hence the plot does not display the tendency for more prominent pitch accents to occur in contrastive focus [e.g., 27] than information focus. However, we observe a majority of accents containing a high accentual tone H*, well in line with studies on prosodic focus in German, e.g., [27]. Roughly, the H* part appears to be related to convey the meaning of adding new information to the common ground, e.g., [28].

In the non-focus condition (left bar), there is a large number of iconic gestures on unaccented referents (red; 75%). The remaining distribution of pitch accents is impressionistically equivalent to distribution in focus, just in a squeezed range. From the comparison of focus and non-focus, it can be seen that focus, unsurprisingly, requires prosodic prominence (presence of a pitch accent).

4.3 Temporal alignment

Figure 3 presents the distribution of temporal alignment of iconic gesture apexes with pitch accents in focus. It can be seen that there is a clear agglomeration of apex-accent alignment around zero with a standard deviation (SD) of 267ms. Zero represents the exact congruence of the pitch accent and the gesture apex. The mean of alignment of 41ms is behind zero. 91.3% of the focused pairs (n=486) were produced within a distance of -500 to 500ms between the two targets. 53% of accents followed the apex.

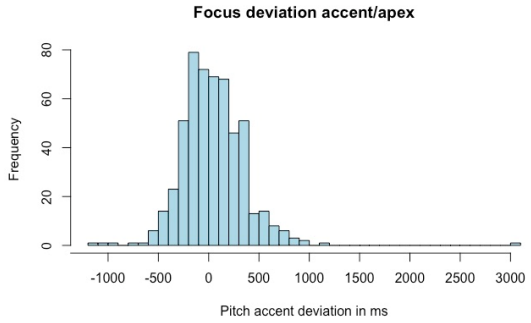

Figure 3: Distribution of the apex-accent distance on focus. - : accent precedes apex; + : accent follows apex.

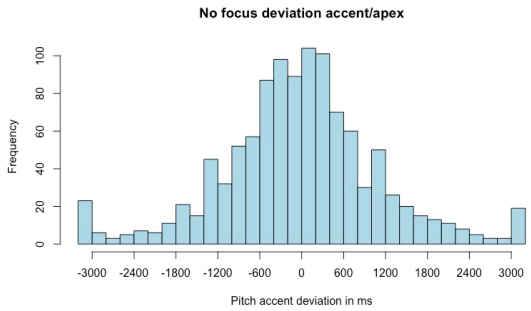

Figure 4: Distribution of the apex-accent distance on non-focus. - : accent precedes apex; + : accent follows apex.

Figure 4 displays the distribution of temporal alignment of iconic gesture apexes with pitch accents in non-focus. The distribution is spread over the full time span, though cumulating around zero. This large deviation of apex-accent alignment amounts in a standard deviation of 416ms. The mean of alignment appears before zero, though with -9ms very close to the midpoint. Overall, comparing with the distribution in focus, the apex-accent alignment in non-focus is less precise than in focus. Only 44.7% of the focused pairs (n=490) were produced within a

distance of -500 to 500ms between the two targets. 49.1% of accents followed the apex.

5. DISCUSSION

The general question of this study was to explore whether and how iconic gestures interact with prosodic prominence, that is whether a pragmatic effect due to focus might be added to the otherwise semantic function of iconic gestures. Starting from a general gesture-prosody link that co-speech gestures are synchronized with prominent syllables or pitch accents [1], we observe that iconic gestures occur in prominent contexts, i.e. in focus (about 25%), but more often in less prominent background contexts (about 40%). If occurring in prominent contexts, iconic gestures always co-occur with pitch accents. In non-focus contexts, the majority of iconic gestures occurs on unaccented elements.

In general, the distribution of iconic gestures in relation to pitch accentuation in focus and non-focus is very similar to that of non-referential gestures [26]. This may point to the fact that focus itself does not impose a preference for a particular type of gesture. Of course, this has to be shown for other gesture types like metaphoric and deictic gestures. Nevertheless, our results point to a similar behaviour of the different gesture types, which seems comparable to the multimodal marking of information status of discourse referents in English [17].

Looking at the distribution of iconic gestures occurring on different pitch accent types, we can observe that although the majority of iconic gestures occurs on unaccented elements in non-focus contexts, the distribution of different pitch accent types accompanied by iconic gestures is otherwise similar in focus and non-focus contexts. Due to the high number of unaccented material, the distribution is squeezed in a smaller range than in focus. Also in this aspect, we observe a similar behaviour of iconic and non-referential gestures [see 26].

The core question of this study was whether the temporal synchronization between iconic gestures and pitch accents is affected by focus. Our results point to a clear difference between focus and non-focus contexts. In focus, iconic gestures are synchronized very tightly with the pitch accent and show less temporal deviation than in non-focus. This effects seem to pattern with a general effect of hyperarticulation in speech [29]. Hence, focus clearly constitutes one of the factors that impacts on the variation in speech along the hypo- and hyperarticulation continuum. In addition, focus more generally affects the multimodal dimension of speech in that both the speech signal and the occurring co-speech gesture together undergo hyperarticulation.

A further conclusion concerns the function of iconic gestures. While iconic gestures are known to express semantic meaning, they also mark prominence. Iconic gestures occur on focused referents to a considerable amount. And these gestures are tightly synchronized with pitch accents. This tight synchrony in focus seems to add a pragmatic function of highlighting to the semantic meaning of iconic gestures. A co-speech gesture can hence express two functions at the same time. Given an identical form of iconic gesture, it visually supports the expressed semantic content of speech [1, 6], and, at the same time, it expresses a discourse function of highlighting information [7–9]. Our findings support the M3D proposal of multiple dimensions to a gesture [4, 17]. There, the idea that a gesture can be labelled at different layers, i.e. form, semantic function and pragmatic function was proposed. Our data show a clear case of support of this differentiation of gestural layers, especially taking into account the similar patterns of iconic and non-referential gestures regarding structural marking.

Future research has to show whether co-speech gestures show a use of a larger space or kinematic movements when expressing two functions at the same time, or whether focus would induce this effect both in discourse structuring non-referential gestures and in referential gestures that express both semantic and pragmatic meaning.

6. CONCLUSION

This study was concerned with the interaction between iconic gestures and prominence. Results show, not surprisingly, that focus requires prosodic marking. In addition, a quarter of those cases were accompanied by iconic gestures. Comparing the synchronization of iconic gestures in focus and non-focus contexts, our results show tighter alignment in focus than in non-focus. This points to two facts. First, focus goes hand in hand with more articulatory effort in speech [29] and prosody [see e.g. 30]. Our results on a closer temporal synchronization of iconic gestures with pitch accent under focus can hence be considered an instance of gestural hyperarticulation. Second, a pragmatic function of highlighting seems to be added to iconic gestures to their otherwise ascribed dimension of expressing a semantic relation to speech concepts.

7. ACKNOWLEDGEMENTS

Thanks to the DFG (KU 2323/5-1), as part of SPP 2392 Visual Communication (ViCom), Frankfurt am Main, Germany, and Goethe University Frankfurt for funding this research. Thanks to Andy Lücking for providing access to the SaGA corpus.

8. REFERENCES

[1] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago: U Chicago Press, 1992.

[2] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser, "The Bielefeld Speech and Gesture Alignment Corpus (SaGA)," in *LREC 2010 Workshop: Multimodal Corpora–Advances in Capturing, Coding and Analyzing Multimodality*, 2010, pp. 92–98.

[3] A. Kendon, *Gesture: Visible action as utterance*. Cambridge: CUP, 2004.

[4] P. L. Rohrer *et al.*, *The MultiModal MultiDimensional (M3D) labeling system for the annotation of audiovisual corpora: Gesture Labeling Manual*. UPF Barcelona, 2020.

[5] E. Krahmer and M. Swerts, "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," *J Memory and Language*, vol. 57, pp. 396–414, 2007.

[6] K. G. Kandana Arachchige, I. Simoes Loureiro, W. Blekic, M. Rossignol, and L. Lefebvre, "The Role of Iconic Gestures in Speech Comprehension: An Overview of Various Methodologies," *Front. Psychol.*, vol. 12, p. 634074, 2021.

[7] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica*, vol. 55, no. 3, pp. 243–276, 2008.

[8] M. Götze *et al.*, "Information structure," in *Interdisciplinary Studies on Information Structure (ISIS)*, vol. 7, S. Dipper, M. Götze, and S. Skopeteas, Eds., Potsdam: Universitätsverlag Potsdam, 2007, pp. 147–187.

[9] M. Zimmermann and E. Onea, "Focus marking and focus interpretation," *Lingua*, vol. 121, no. 11, pp. 1651–1670, 2011.

[10] F. Kügler and S. Calhoun, "Prosodic Encoding of Information Structure: A typological perspective," in *The Oxford Handbook of Language Prosody*, C. Gussenhoven and A. Chen, Eds., Oxford: Oxford University Press, 2020, pp. 453–467.

[11] M. Nespor and I. Vogel, *Prosodic phonology*. Dordrecht: Foris Publ., 1986.

[12] E. Selkirk, "On derived domains in sentence phonology," *Phonology Yearbook*, vol. 3, pp. 371–405, 1986.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: CUP, 2004.

[14] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed., Oxford: Oxford University Press, 2005, pp. 55–83.

[15] E. O. Selkirk, "The syntax-phonology interface," in *The Handbook of Phonological Theory*, J. A. Goldsmith, J. Riggle, and A. C. L. Yu, Eds., 2nd ed., Chichester: Wiley-Blackwell, 2011, pp. 435–484.

[16] S. Baumann and C. T. Röhr, "The perceptual prominence of pitch accent types in German," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 2015, p. 384.

[17] P. L. Rohrer, "A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system," PhD thesis, Universitat Pompeu Fabra, Barcelona, 2022.

[18] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, M. R. Key, Ed., Berlin: Mouton, 1980, pp. 207–227.

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

[20] N. Esteve-Gibert and P. Prieto, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *J Speech Lang Hear Res*, vol. 56, no. 3, pp. 850–864, 2013.

[21] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, pp. 71–89, 2012.

[22] S. Im and S. Baumann, "Probabilistic relation between co-speech gestures, pitch accents and information status," *Proceedings of the Linguistic Society of America*, vol. 5, no. 1, pp. 685–697, 2020.

[23] D. McNeill, "Gesture: a psycholinguistic approach," in *Encyclopedia of language and linguistics*, E. K. Brown, Ed., 2nd ed., Amsterdam: Elsevier, 2006, pp. 58–66.

[24] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]: Version 6.3.03 (http://www.praat.org/)*, 2022.

[25] S. Baumann, M. Grice, and S. Steindamm, "Prosodic Marking of Focus Domains - Categorical or Gradient?," in *Proceedings of Speech Prosody 2006, Dresden, Germany*, Dresden, 2006, pp. 301–304.

[26] A. Gregori, "Co-speech Gestures, Information Structure and Prosody: A Corpus Study on Prominence Peak Alignment," MA thesis, Goethe Universität Frankfurt, Frankfurt am Main, 2022.

[27] M. Grice, S. Baumann, and N. Jagdfeld, "Tonal association and derived nuclear accents-The case of downstepping contours in German," *Lingua*, vol. 119, pp. 881–905, 2009.

[28] J. Peters, *Intonation*. Heidelberg: Universitätsverlag Winter, 2021.

[29] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Kluwer, 1990, pp. 403–439.

[30] J. Hanssen, J. Peters, and C. Gussenhoven, "Prosodic Effects of Focus in Dutch Declaratives," in *Proceedings of Speech Prosody 2008 Conference*, 2008, pp. 609–612.