Pitch Accent Trajectories across Different Conditions of Visibility and Information Structure — Evidence from Spontaneous Dyadic Interaction

Petra Wagner, Nataliya Bryhadyr, Marin Schröer

Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies
Bielefeld University, Germany

{petra.wagner, n.bryhadyr, marin.schroeer}@uni-bielefeld.de

Abstract

Previous research identified a differential contribution of information structure and the visibility of facial and contextual information to the acoustic-prosodic expression of pitch accents. However, it is unclear whether pitch accent shapes are affected by these conditions as well. To investigate whether varying context cues have a differentiated impact on pitch accent trajectories produced in conversational interaction, we modified the visibility conditions in a spontaneous dyadic interaction task, i.e. a verbalized version of TicTacToe. Besides varying visibility, the game task allows for measuring the impact of information-structure on pitch accent trajectories, differentiating important and unpredictable game moves. Using GAMMs on four speaker groups (identified by a cluster analysis), we could isolate varying strategies of prosodic adaptation to contextual change. While few speaker groups showed a reaction to the availability of visible context cues (facial prosody or executed game moves), all groups differentiated the verbalization of unpredictable and predictable game moves with a group-specific trajectory adaptation. The importance of game moves resulted in differentiated adaptations in two out of four speaker groups. The detected strategic trajectory adaptations were characterized by different characteristics of boundary tones, adaptations of the global f0-level, or the shape of the corresponding pitch accent.

Index Terms: dialogue, prosody, pitch accents, visibility

1. Introduction

It is well described that contextually novel, surprising, important, contrastive or somehow discourse-relevant words are made prosodically prominent across a number of typologically diverse languages [1, 2, 3, 4]. What is less well understood is whether these various prominence-cueing sources should be modeled as a single one-dimensional concept of "information structure", or whether the different functions of prominence are phonologically differentiated, e.g. into a notion of "contrast" and "novelty" [5]. It is also largely unclear how pragmatic functions interact with the expression of paralinguistic notions [6]. Watson et al. [7] investigated whether different types of information-structure trigger different types of prosodic prominence. They operationalized the difference between *importance accents* and *unpredictability accents* by measuring verbalized game moves in games of TicTacToe. In early stages of the game, the moves are relatively unpredictable, but also less relevant, as they are not decisive for the outcome of the game. Later on, the game moves are highly predictable, but relevant, as they typically prevent the interlocutor from winning, or may constitute winning moves (cf. Figure 1). Hence, while there are information-structural reasons for accenting each verbalized move, the type of accentuation may vary as a function of the accent type, i.e., accents for *importance*, or accents for *unpredictability*. For American English, [7] found that accents expressing unpredictability are longer and are produced with a higher F0 excursion, while accents related to relevance are louder. For German, [8] found that unpredictable accents are longer than important (and predictable) ones, but show no differentiated effect on loudness, intensity, pitch excursion or height.

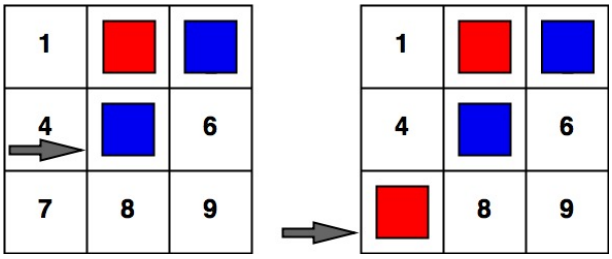

Figure 1: *Examples for an unpredictable (and unimportant) move (left), and an important (and predictable) move (right).*

However, these analyses failed to look at pitch accent shapes, which have shown to contribute to the perception of prosodic prominence in German, with low accents being less prominent than high accents, and late accents being more prominent than early ones [9]. Also, prosodic phonology has often attributed different functions to pitch accent shapes: [10] finds pitch accents with an early peak to express contextual givenness (or predictability), middle peaks to express novelty (or unpredictability), and late peaks to express paralinguistic emphasis (or importance). Along similar lines, [11] find that less prominent pitch accents with falling shapes and low valleys (H−L*) or deaccentuation express contextual accessability (or predictability). [12] confirm these tendencies also for spontaneous productions, while [13] find that speakers have individual strategies for expressing prosodic contrasts. To our knowledge, no study has tried to disentangle different levels of predictability (roughly corresponding to contextual accessability) and the more paralinguistic notion of importance with respect to the f0 trajectories of pitch accents. An investigation of these constitutes the first goal of this paper, using Generalized Additive Mixed Models (GAMMs), as these allow for a fine-grained analysis of nonlinear shapes over the course of time, without the need for a prior manual assessment of contours.

A second goal of our study relates to the impact of visible access to contextual information on the f0 trajectories. From [8], we know that access to facial expressions leads to a slight increase in mean f0, while visible access to the content being verbalized leads to a slight increase in speech tempo. The latter finding is in line with information theoretical predictions, as vis-

ible access to the expressed context renders a message largely predictable, resulting in a less prominent production. The former result goes against our expectations, as the visible access to facial expression should lead to a better transmission channel, and should result in less production effort [14]. We therefore conclude that the slight overall increase in f0 is the result of a stronger experienced co-presence, resulting in more engagement [15].

In line with the prior research, we formulate the following hypotheses: (1) We expect unpredictable accents to be realized with prominent, high, rather than low pitch accents, and possibly with later peaks. (2) Due to our setting contrasting unpredictability and importance (cf. below), important accents could be either signaled as being accessible (i.e., with an early peak and a low accent), or emphatic (i.e., with a late peak). (3) In line with information structural and speech optimization theories, we furthermore expect a decrease in the expression of prosodic prominence given visibility, resulting in low accents or early accent peaks.

2. Methods

2.1. Data

All analyses are carried out on a data set described in full detail in [8]. The data contains 20 spontaneous, task-oriented dialogues of 40 native speakers of German (equal social status, same or mixed gender). Data from one speaker was discarded for technical reasons. The interlocutors were engaged in a verbalized version of TicTacToe, describing their moves on a shared vertical 3x3 grid, while simultaneously playing the game, marking the moves using colored felt squares (cf. Figures 1, 2). The fields in the TicTacToe grid are numbered from $1-9$, and the players use these fields to unambiguously refer to their ongoing move. The participants did not receive instructions as to how they should formulate their game moves, but the vast majority of them was realized in the form of short utterances, where the target move (e.g. the named digit corresponding to the field on the grid) is the last word in the utterance, coinciding with a nuclear pitch accent and a prosodic boundary:

`Mein nächster Zug geht auf FÜNF.`
(Engl.: My next move goes on FIVE.)

We analysed the prosodic realizations of these verbalized target moves ("1-9") in utterance final position.

2.2. Study design

Each dyad performed 4 games of TicTacToe with 4 different visibility conditions, which served as independent variables: (1) Full visibility, (2) facial visibility, where the game moves are not visible due to an intransparent game board, (3) facial visibility, where the game moves are visible through a transparent game board, but where access to facial expressions is obstructed by a light but opaque curtain, and (4) No visibility, where both facial expressions and game moves are not visible to interlocutors (cf. Figure 2).

We used the TicTacToe Setting to disentangle two aspects of information structure, namely *predictability* and *importance*, which also served as independent variables. The initial moves were (quasi-randomly) predefined by the experimenter and only repeated by the participant. These were defined as fully predictable ("9"). The second move was annotated as least predictable ("1"), and the next moves were defined with increasing predictability in course of the game. For statistical analyses, moves with a predictability $> 4 =$ were coded as "predictable", and remaining moves as "unpredictable". Importance was operationalized in a binary fashion, with moves that prevent or constitute a winning move being annotated as "important", all others as "unimportant" (cf. 1). Due to the design, important moves tend to be predictable and vice versa. An overview of the study design is illustrated in Table 1.

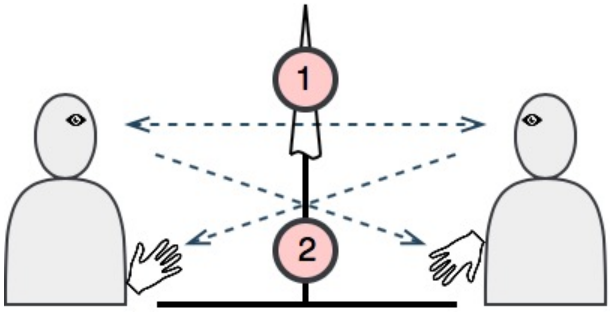

Figure 2: *Recording setup, with (1) being a light, opaque curtain and (2) being a vertical game board. Removal of the curtain allows for mutual facial visibility, and a transparent game board for interlocutors' manual visibility*

visibility			information structure	
setting	facial	manual	important	unpredictable
full visibility	+	+	+ for decisive moves, else -	+ for early moves, else -
visible face	+	-		
visible hands	-	+		
no visibility	-	-		

Table 1: *Controlled independent variables in the study design.*

2.3. Annotations and data extraction

In each dyad, the verbalizations of the game move targets (verbalizations of "1-9") as well as the corresponding move's relevance and predictability were annotated manually using Praat [16]. We restricted our analysis to utterance final realizations, almost always coinciding with a (nuclear) accent and a following boundary tone. Using a Praat script, we extracted f0 contours for the target utterances, using the Praat built-in pitch tracking function: For each target move, F0 trajectories were calculated in time windows of 100ms, z-scored by speaker, and the resulting contours were time-normalized between 0 and 1 for each target utterance. The resulting contours then served as dependent variables in the subsequent analyses using Generative Additive Mixed Models (GAMMs) implemented in the R-packages `mgcv` (vers. 1.8.22) and `itsadug` (vers. 2.3.2) [17, 18], and closely following the methodological suggestions in [19]. To determine areas of significant difference between f0 shapes, we calculated the differences between the smooths corresponding to different conditions, and determined those intervals where the confidence bands are different from zero.

2.4. GAMM analyses on pooled data

In a first step, GAMMS were built for the pooled data, i.e., for the f0 trajectories of the target verbalizations of each game

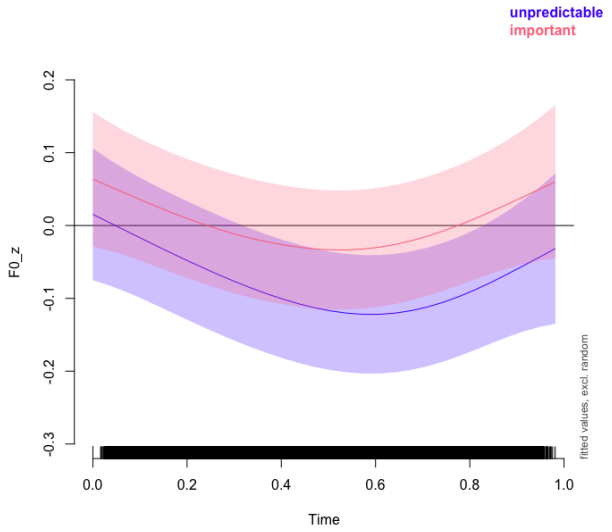

Figure 3: *Non-linear smooths for f0 trajectories in important and predictable words (pooled data). Shaded bands show pointwise 95% confidence intervals.*

move serving as dependent variables. We built separate models for testing the impact of different levels of the fixed factors *predictability*, *importance*, *facial visibility* and *manual visibility*, to test their impact on the potentially non-linear f0 trajectories over time. We used thin place regression splines as smooth factors, and checked for oversmoothing. As random factors, we added random intercepts for *speaker* and *word* (i.e., the target of the game move uttered, corresponding to the numbers 1–9), as model comparisons did not show significant differences for more complex models including random slopes or non-linear random effects.

2.5. GAMM analyses on clustered data

Typically, prosodic analyses show a large degree of within and between-speaker variation. We therefore wanted to find out whether there are group-specific strategies for reacting to different conditions of information structure or visibility. To this end, we first performed a cluster analysis on speaker-wise cross-correlations on the mean f0 levels across the different conditions of information structure and visibility present in the data (cf. Figure 3). The cluster analysis was performed with the hierarchical clustering method using `hclust` function implemented in the `stats` package (version 3.4.3) within R [20]. We then split the f0 trajectories data according to the results of the first four clusters detected, resulting in one group of 4 speakers (cl1), a second group of 15 speakers (cl2), a third group of 13 speakers (cl3), and a forth group of 9 speakers (cl4). For each of these clusters, we subsequently performed the same GAMM-based analyses as were performed on the pooled data set.

3. Results

An overview of significant differences in f0 trajectories across the various contrasts is presented in Table 2

3.1. Pooled data

For the pooled data, only unpredictability had a significant effect on f0 trajectories, mainly affecting the global height of the f0, but in contrast to our expectations, predictable words are showing a somewhat higher global f0 contour than unpredictable ones. We furthermore found a global significant difference when contrasting unpredictable and important accents, with important accents showing a slightly higher f0 contour (c.f., Figure 3). For all other independent variables, no significant impacts on f0 trajectories were detectable.

3.2. Clustered data for individual speaker groups

All clusters showed an independent significant impact of unpredictability on f0 trajectories, mostly by a clear differences in contour shapes: unpredictable words are expressed with a falling contour, predictable ones with a falling-rising shape. Importance is marked in two speaker groups (cl1, cl4), with both showing falling-rising contours for important, and falling contours for unimportant accents (cf. Figure 4. However, all 4 clusters significantly contrast unpredictable and important pitch accents, but in different ways: while all clusters produce important accents at a higher global f0 level, cl1 prefers falling contours for both important and unpredictable accents, Cl2 prefer a falling pattern for important, and a fall-rise pattern for unpredictable ones, cl3 show fall-rise patterns for both (at different f0 levels), and cl4 show falling patterns for unpredictable, and flat patterns for important ones.

For the different visibility conditions, only single clusters showed a significant influence: Cl1 preferred falling contours if facial visibility was given, and rising contours if interlocutors could not see each others faces (cf. Figure 5). If interlocutors had visible information about the game moves (manual visibility), cl3 showed comparatively lower accent valleys (cf. Figure 6).

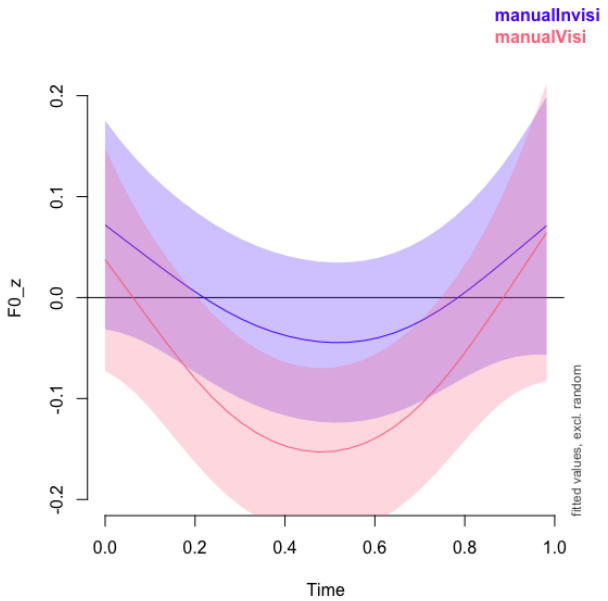

Figure 6: *Non-linear smooths show lower accent valleys if hands are visible for speaker group cl3. Shaded bands show pointwise 95% confidence intervals.*

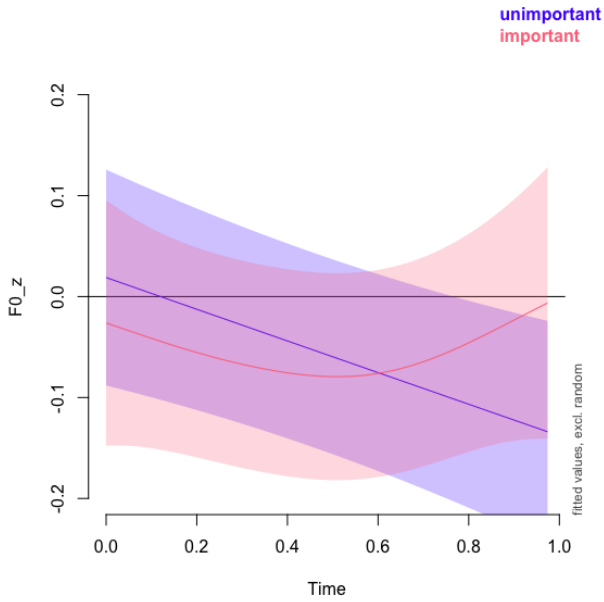

Figure 4: *Non-linear smooths show fall-rise contours in important and falling contours in unimportant word accents for speaker group cl4. Shaded bands show pointwise 95% confidence intervals.*

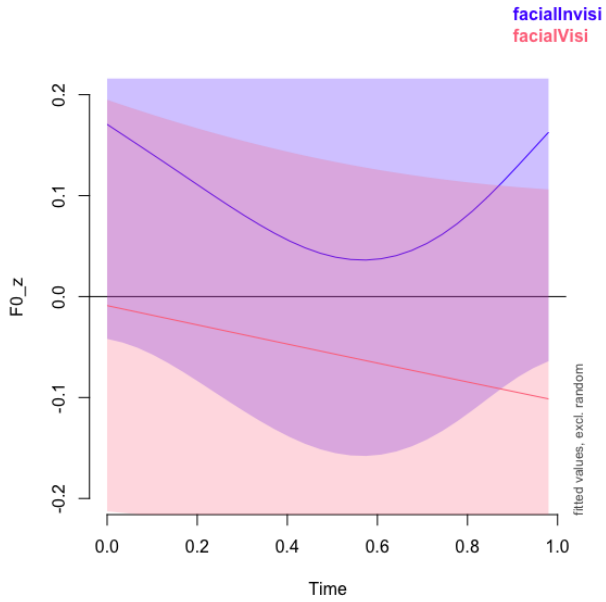

Figure 5: *Non-linear smooths show falling accents if faces are visible, rising ones if faces are not visible for speaker group cl1. Shaded bands show pointwise 95% confidence intervals.*

group	contrasted factors				
	import. vs. unpred.	unpred. vs. other	import. vs. other	fac. visib. vs. other	man. visib. vs. other
pooled	✓	✓			
cl1	✓	✓	✓	✓	
cl2	✓	✓			✓
cl3	✓	✓			
cl4	✓	✓	✓		

Table 2: *Overview of significant contrasts between smooths describing f0 trajectories.*

4. Discussion

Our GAMM-based analysis allowed for a first insight of a differentiated impact of information-structural and visibility conditions on the realization of f0 trajectories in pitch accents.

Generally, the lack of significant results for the pooled data is a sign of the high amount of individual variation in how f0 trajectories are employed in signalling pragmatic and paralinguistic functions. When pooling the data across participants, the only stable effects are the differentiation of lower unpredictable and higher predictable as well as lower unpredictable and higher important accents. The former, surprising result is understood much better when looking at the contour shapes for the individual clusters, showing a general trend of falling unpredictable vs. falling-rising (predictable) accents. While this result is contrary to our hypothesis, it may be a task related effect, where the falling-rising pattern asks for quick continuation (after a "boring" predictable event), while unpredictable accents may rather indicate a termination of a surprising game move, which may need some time to react to.

The global contrast between important and unpredictable events indicates that in terms of pitch, more effort (and probably more prominence) is invested in the pronunciation of important events. This contradicts our hypothesis, as we expected unpredictable accents to be produced with higher peaks. Also, we could not isolate a global preference for distinguishing pitch shapes, but rather found a high degree of variation. Obviously, the signaling of importance "beats" unpredictability in terms of prosodic effort invested as f0, as this is the case for all speaker groups. This result is intriguing, as despite the fact that the dialogue task was competitive, there was no risk involved, and the task was quite easy. Obviously, it was still sufficient to initiate an atmosphere of engagement.

Importance only was only differentiated by two groups of speakers, showing a falling contour for unimportant, and a falling-rising contour for important events. This might possibly coincide with an early pitch accent, followed by a rising boundary tone, or a late pitch accent, and is in line with our hypotheses. As we also marked those moves as important, which prevented the opponent from winning, a high boundary may be used as a cue to invite a metaphorical "answer" on the game board to a clever move, and may be indicative of high engagement. At this point, the result is difficult to interpret from a paralinguistic and pragmatic point of view. However, only a small subset of speakers used this strategy.

Similarly, only a subset of our speakers reacted to the presence or absence of the interlocutor's visibility. However, if they did, their behaviour corresponded with our expectations: A lack of facial visibility resulted in more pronounced accentual shapes and high boundary tones or late accents, indicating a stronger prosodic prominence. A lack of manual visibility resulted in a higher f0 level, expressing a higher degree of invested prosodic effort. These findings are in line with effort optimization models of speech production such as Lindblom's H&H theory [14].

5. References

[1] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgan, and M. Pollack, Eds. Cambridge MA: MIT Press, 1990, pp. 271–311.

[2] Y. Xu, "Effects of tone and focus on the formation and alignment of f0 contours," *Journal of Phonetics*, no. 27, pp. 55– 105, 1999.

[3] C. Féry and F. Kügler, "Pitch accent scaling on given, new and focused constituents in German," *Journal of Phonetics*, vol. 36, pp. 680 – 703, 2008.

[4] S. Skopeteas and C. Féry, "Effect of narrow focus on tonal realization in Georgian." in *Proceedings of Speech Prosody 2010*, Chicago, Illinois, 2010.

[5] A. Riester and S. Baumann, "Focus triggers and focus types from a corpus perspective," *Dialogue and Discourse*, vol. 4, no. 2, pp. 215–248, 2013.

[6] M. Grice and S. Baumann, "Deutsche intonation und GToBI," *Linguistische Berichte*, vol. 191, pp. 267–298, 2002.

[7] D. Watson, J. Arnold, and M. K. Tanenhaus, "Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production," *Cognition*, vol. 106, no. 3, pp. 1548–1557, 2008.

[8] P. Wagner, N. Bryhadyr, and M. Schröer, "Does information structural prosody change under different visibility conditions?" in *Proceedings of the International Congress of the Phonetic Sciences*, Melbourne, Australia, 2019.

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

[10] S. Baumann and M. Grice, "Terminal intonation patterns in single-accent utterances of german: phonetics, phonology and semantics," *AIPUK*, vol. 25, p. 115185, 1991.

[11] ——, "The intonation of accessibility," *Journal of Pragmatics*, vol. 38, pp. 1636–1657, 2006.

[12] S. Baumann and A. Riester, "Coreference, lexical givenness and prosody in German," *Lingua*, vol. 136, pp. 16–37, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.lingua.2013.07.012

[13] O. Niebuhr, M. D'Imperio, B. Gili Fivela, and F. Cangemi, "Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category," in *International Congress of Phonetic Sciences 17*, Hong Kong, China, Aug. 2011, pp. 120–123. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01514844

[14] B. Lindblom, *Explaining Phonetic Variation: A Sketch of the H&H Theory*. Kluwer Academic Publishers, 1990, pp. 403–439.

[15] S. Shahid, E. Krahmer, and M. Swerts, "Alone or together: Exploring the effect of physical co-presence on the emotional expressions of game playing children across cultures," in *Fun and Games*, P. Markopoulos, B. de Ruyter, W. IJsselsteijn, and D. Rowland, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 94–105.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.0.49," 2019. [Online]. Available: http://www.praat.org/

[17] S. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[18] J. van Rij, M. Wieling, R. H. Baayen, and H. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2017, r package version 2.3.

[19] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between l1 and l2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/