What You See is What You Get Prosodically Less — Visibility Shapes Prosodic Prominence Production in Spontaneous Interaction

Petra Wagner[1,2], *Nataliya Bryhadyr*[1]

[1]Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[2]Center of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University, Germany
{petra.wagner,n.bryhadyr}@uni-bielefeld.de

Abstract

We investigated the expression of prosodic prominence related to unpredictability and relevance in spontaneous dyadic interactions in which interlocutors could or could not see each other. Interactions between visibility and prominence were analyzed in a verbal version of the game TicTacToe. This setting allows for disentangling different types of information structure: early moves tend to be unpredictable, but are typically irrelevant for the immediate outcome of the game, while late moves tend to be predictable but relevant, as they usually prevent an opponent's winning move or constitute a winning move by themselves.

Our analyses on German reveal that prominence expression is affected globally by visibility conditions: speech becomes overall softer and faster when interlocutors can see each other. However, speakers differentiate unpredictability and relevance-related accents rather consistently using intensity cues both under visibility and invisibility conditions. We also find that pitch excursions related to prosodic information structure are not affected by visibility. Our findings support effort-optimization models of speech production, but also models that regard speech production as an integrated bimodal process with a high degree of congruency across domains.

1. Introduction

As many of our everyday interactions take place both under visibility conditions (e.g. face-to-face dialogues) and invisibility conditions (e.g. telephone conversations), it is of interest to see whether and how our prosodic expressions are shaped by the factor of visibility. Much research on the prosodic realization of information structure has shown that contextually novel, surprising, important, contrastive or somehow discourse-relevant words are made prosodically prominent, e.g. using pitch accentuation or lengthening, across a number of typologically diverse languages [1, 2, 3, 4]. However, it is still a matter of debate, whether information structure should be treated as a one-dimensional concept, or whether one needs to take into account different types of "focus accent triggers" such as novelty/givenness vs. contrast/focus to fully understand it [5]. Furthermore, we do not yet know to what extent information that is accessible via a visual channel alters the prosodic expression of the same information in the acoustic channel.

Watson et al. [6] investigated whether different types of information-structure trigger different types of prosodic prominence in American English. They operationalized the difference between *relevance accents* (roughly corresponding to "focus" in information theory) and *unpredictability accents* (roughly corresponding to "new" in information theory)) by measuring different types of game moves in a verbal version of TicTacToe. In early stages of the game, the moves are relatively unpredictable, hence tend to be accented, but also less relevant, as they are not decisive for the outcome of the game. Later on, the game moves are highly predictable, but important, as they typically prevent the interlocutor from winning, or may constitute winning moves (cf. Figure 1). Although highly predictable in nature, these late moves are likely to be produced prominently on the basis of their relevance for the outcome of the game. For American English, [6] found a difference in the prosodic realizations of these two types of prominence: accents expressing unpredictability are longer in duration and are produced with a higher F0 excursion, while accents related to relevance are louder.

In their study, [6] controlled for visibility, i.e. the interlocutors did not see each other during their interaction. The reason for this control was probably due to the circumstance that the verbalized moves are entirely redundant under visibility conditions, i.e. there is no further need to express a game move's relevance or unpredictability when this information is fully shared between the interlocutors. Furthermore, visibility of the interlocutors' head and facial movements may enhance intelligibility and prominence perception [7, 8, 9]. An effort-optimization model (e.g. [10, 11]) of speech production would therefore predict less articulatory effort under visibility conditions. This may result in a deletion of the fine-grained function-specific ways of prominence production detected in American English.

However, research on speech-gesture interaction has shown that speakers prefer a high degree of congruency (or redundancy) between information transported in the visual and in the acoustic channel, thereby enhancing the robustness of verbal communication [12]. This congruency between speech and co-speech gesture appears to be particularly strong for the prosodic domain [13, 14, 15].

Taken together, these findings make no clear predictions about the influence of visibility on prosodic information structure marking: A model focusing on optimizing production effort predicts a decrease in the acoustic prominence marking of information structure under visibility. A model that regards co-

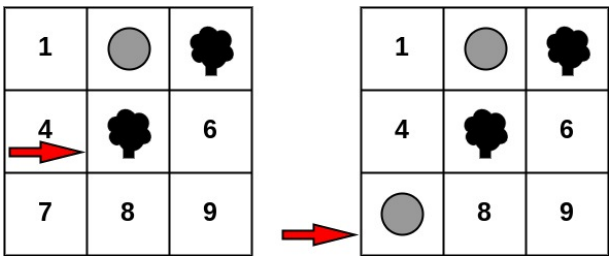

Figure 1: *An unpredictable move on field "5" (left), followed by a relevant move on field "7" (right) in TicTacToe.*

speech movements as a means to support the robustness in information exchange does not, as it relies on congruency between visual and verbal information transmission. Our research thus aims at a better understanding of the following questions:

1. Is the acoustic-prosodic differentiation between unpredictability and relevance found for American English evident in German prosody as well?
2. Is prosodic expression enhanced if interlocutors cannot see each other while interacting, thus supporting effort-minimization accounts of speech production?
3. Is a functional distinction between accents sustained under visibility conditions, when the prosodic information is redundant, supporting integrated bimodal accounts of speech production?

To answer these questions, we replicated Watson et al.'s [6] study with German speakers, but added a recording condition with full visibility between interlocutors. Given the typological similarity of German and American English, we hypothesize that German shows a similar prosodic distinction between *unpredictability accents* and *relevance accents*, especially in the invisibility condition.

In line with theories of speech production that economize production effort, we furthermore hypothesize that visibility leads to an overall reduction in prosodic effort, i.e. we expect speech production to be less loud, faster and having less strong pitch excursions. We expect that this reduction in prosodic effort may delete any fine-grained function-specific prosodic realizations, challenging models that expect a maximal congruence between verbalizations and information transmitted by co-verbal movements.

2. Methods

2.1. Participants

We recorded 20 native speakers of German (10 dyads) engaged in a verbalized version of TicTacToe. The speakers in each dyad were familiar with one another and of equal social status (typically friends). This was done in order to ease the atmosphere (to break the ice) and to encourage an informal, spontaneous speaking style. The participants were not controlled for gender, i.e. dyads had speakers with identical or different gender. With one exception, all speakers were in their twenties. All participants were unpaid volunteers.

2.2. Recording setup

Each dyad was recorded at our faculty's recording studio using Sennheiser neckband microphones in two different recording conditions:

1. Visibility Condition: The players were seated facing each other, with a shared TicTacToe game board placed in the middle (cf. Figure 2).
2. Invisibility Condition: The players were seated on separate tables and were parted from each other by the movable wall, each of them having his or her own game board (cf. Figure 3).

Each player received a set of cut outs in the form of a *'tree'* (ger. *'Baum'*) and a *'ball'* (ger. *'Ball'*) to mark their moves. These forms were chosen as they are easy to depict and trigger voicing. Four games were played per setting. To control for order effects, we used an alternating initial recording condition (visibility or non-visibility) with each newly recorded dyad.

Figure 2: *The recording setting within the visibility condition. The players have shared access to game-related information.*

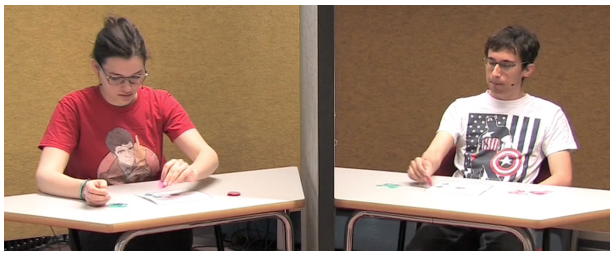

Figure 3: *The recording setting within the invisibility condition. The players have to negotiate all game-related information verbally.*

The game board looked like a normal TicTacToe grid, however with every cell being numbered. This was introduced in order to enable the interlocutors to unambiguously refer to the different cells on the game board using the digits $1-9$. That way, a typical verbalized move is produced by placing a sentence or nuclear accent on the target of the move, which corresponds to one of the numbers available on the game board – these accented verbalizations of numbers are later analyzed for their prosodic realization, e.g.

`Ich lege einen Baum auf Feld FÜNF.`
(engl.: I put a tree on field FIVE.)

In order to prevent the interlocutors from repeating an identical pattern or strategy, prior to each game, the players were informed about a preset first move. Also, the players alternated in setting the first move.

On average, each recording session lasted 8.57 minutes per dyad, resulting in roughly 1.5 hours of recorded speech in total. In addition to the audio data, video recordings were collected which are not part of the analyses reported here.

2.3. Annotations

The verbalized target moves, i.e. the nuclear accented number realizations, were manually annotated using Praat [16] for further acoustic analysis. As the first move was preset as part of the recording setting and made known to both players before being verbalized, it is annotated as neither important nor relevant, i.e. *given*. Please notice that even in these "given" cases, an accent on the target word was perceivable. The remaining moves were annotated for the complementary features *importance* and *relevance*. Due to our restricted game setting, a relevant move is automatically predictable and vice versa (cf. Figure 1). Moves were labeled as relevant when they led to a win (i.e. the move realizes a sequence of three vertical, horizontal, or diagonal moves), or if they blocked a potential winning

move of the game opponent. The remaining moves were considered to be unpredictable. In the case of a tie, the last move was annotated as *given*, as it is fully predictable and irrelevant for the outcome of the game. This leaves us with three distinct types of moves, namely (1) given (predictable & irrelevant), (2) unpredictable, and (3) relevant.

2.4. Acoustic analyses

In order to get results that are comparable with [6], we used similar acoustic features to analyze the prosodic realizations of the target words within the recordings. Please notice that all examined accents are produced sentence finally, thus, will all be equally affected by prosodic boundaries:

- duration (ms)
- intensity (RMS)
- mean F0 (st rel 1 Hz)
- F0-range (st)

All analyzes were performed using the built-in Praat functions with standard settings. In some cases, F0-analyses yielded none or no meaningful values, probably due to heavy glottalization or other artifacts caused by the pitch tracking algorithm. These data points were deleted from further analyses.

3. Results

The dyads without visibility were considerably longer (mean duration $= 4.54$ minutes) compared to the recordings with visibility (mean duration $= 4.04$ minutes), indicating a global effect of visibility on overall speech tempo and task completion time. However, this is probably partly caused by participants having to carry out all game moves by themselves in the invisibility condition.

The data collected in the recordings and subsequent annotations and acoustic measurements were further analyzed with the help of Linear Mixed Models using R (Version 3.1.2) [17] together with the R-packages lme4 (Version 1.1-7) and lmerTest (2.0-25). As we wanted to focus on the contrast between unpredictability and relevance rather than givenness, we performed the analyses on a subset of the data, disregarding given moves. For illustration purposes, we left the measurements for given words in the presented interaction plots. The resulting models contained the 2-level factors *visibility (visible—invisible)* and *accent type (unpredictable—relevant)* as fixed factors and *word* and *participant* as random factors with random intercepts. We also tested for interactions between the factors *visibility* and *accent type*. The various acoustic measurements *duration, intensity, mean F0, F0-range* served as dependent variables.

Each dependent acoustic variable was analyzed by reducing a maximal model in a stepwise fashion by removing all non-significant main effects and interactions through log-likelihood ratio comparisons.

3.1. Duration

A visual analysis of the duration patterns suggests an interaction between visibility and a usage of lengthening that differentiates between unpredictable and relevant moves (cf. Figure 4), with relevant moves being considerably longer than unpredictable ones under invisibility conditions only. However, the log-likelihood comparison of the full model including an interaction and the one without shows a marginally significant effect only ($\chi^2(1) = 2.79, p = 0.095$) and is no longer pursued as the model comparison also fails to find a significant effect of accent type. The reduced final model contains the fixed factor *visibility* only.

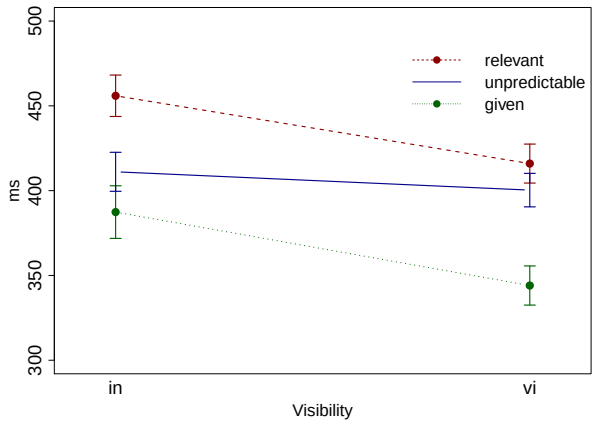

Figure 4: *Mean durations for relevant, unpredictable and given moves under visibility and invisibility conditions.*

Presence or absence of visibility had a significant effect on duration: when interlocutors could see each other, they produced their moves significantly ($t(507) = -3.6, p < 0.0001$) faster ($-32ms$, $SD = 8.8ms$).

3.2. Intensity

A visual analysis of the intensity patterns suggests no interactions of visibility and accent type (cf. Figure 5). This is supported by a log-likelihood comparison.

The presence or absence of visibility had a significant effect on intensity: when interlocutors could see each other, they produced their target moves significantly ($t(504.9) = -2.18, p = 0.0029$) softer ($-4.7$, $SD = 2.2$).

Furthermore, Figure 5 hints at a usage of intensity that differentiates between unpredictable and relevant moves, with unpredictable moves being louder than relevant ones both under visibility and invisibility conditions. This impression is supported, with unpredictable moves being realized significantly ($t(518.2) = 2.1, p = 0.034$) louder ($+4.9$, $SD = 2.3$) than relevant ones.

3.3. Pitch

A visual analysis of the behavior of mean pitch suggests neither an effect of visibility nor a functional prosodic differentiation between relevant and unpredictable accents (cf. Figure 6).

These impressions are confirmed by the statistical analyses, yielding neither an influence of visibility nor of accent type on mean pitch. The dominant factor influencing pitch excursion appears to be the generally high information status of a word, that distinguishes it from given referents. This assumption receives support from a model comparison where words annotated as contextually given were included. Here, both unpredictable ($t(648.1) = 3.4, p = 0.0007$) and relevant ($t(653.7) = 3.4, p = 0.0005$) accents make a significant contribution and lead to a very similar increase in mean F0 of roughly 2.5 semitones (unpredictability: $2.5, SD = 0.75$; relevance: $2.6, SD = 0.73$).

The analysis of pitch range yielded no significant results.

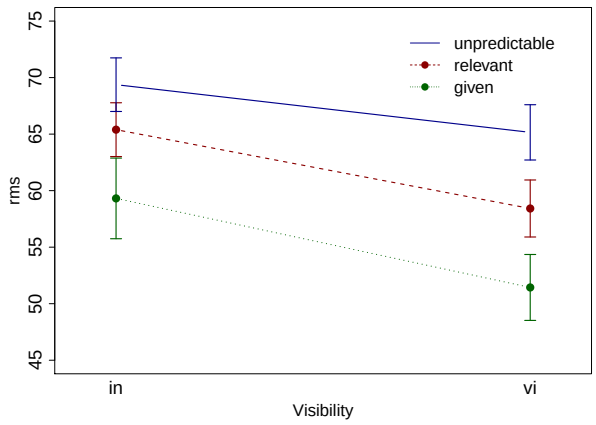

Figure 5: *Mean intensities (RMS) for relevant, unpredictable and given moves under visibility and invisibility conditions.*

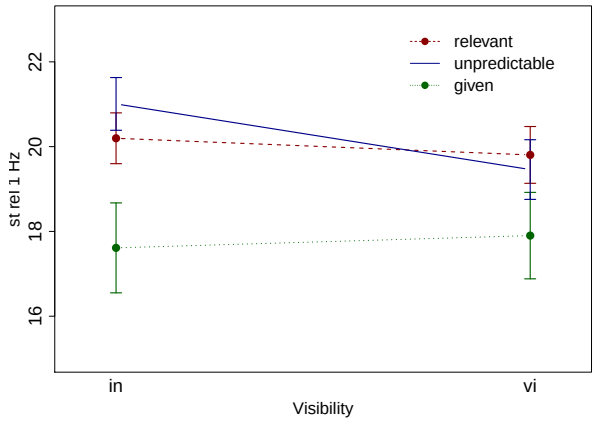

Figure 6: *Mean F0 (st rel 1 Hz) for relevant, unpredictable and given moves under visibility and invisibility conditions.*

4. Discussion

Our first question was whether German speakers make the same functional differentiation between *relevance accents* and *unpredictability accents* as speakers of American English. While we did find that German speakers systematically differentiate between these two accents, they do so using different prosodic means: unpredictable accents are louder than relevant ones, but unlike in American English, there exists no difference in duration or fundamental frequency excursion. Also, German uses intensity in the opposite direction compared to American English, where accents related to relevance were significantly louder than unpredictable ones. Given the typological proximity of American English and German, this result may come as a surprise, but is easy to interpret in the light of findings that prosodic variation may be affected by geographical as well as typological neighborhood [18]. Given the geographical distance of Germany and the US, an analysis of German or other Germanic languages spoken in the US would provide material for an interesting follow-up study. The null result on pitch calls for a deeper re-analysis, as our measurements although chosen to faithfully replicate [6] should be taken with a grain of salt, as they did not take into account fundamental frequency movement or pitch peak alignment with the accented syllable, which have shown to be influential in the expression of different levels of prosodic prominence in German accents [19].

Our second question related to a potential impact of visibility on prosodic expression. As we indeed found an influence on overall interaction time (dialogues are shorter under visibility), word duration (accented words are shorter under visibility), and intensity (accented words are softer under visibility), effort-optimization accounts of speech production receive support. Evidently, speakers invest less prosodic effort when their productions can be seen as well as heard by their interlocutors, who may rely on visual cues to sustain intelligibility. However, the lack of "prosodic reduction" in pitch excursion points into a different direction. Given less available duration to produce pitch accents, speakers invest more rather than less effort in their pitch accent production when their interlocutors can see as well as hear them. This illustrates that speech effort minimization may not go "all the way", and speakers take care that pitch accent function is not jeopardized, even if the information that is transmitted verbally is redundant. This finding supports an integrated view of speech-gesture production, where speech and co-speech gesture typically transport congruent information.

Lastly, we wanted to find out whether a functional differentiation between *unpredictability accents* and *relevance accents* as studied by [6] under invisibility extends to visibility settings, despite potential impacts of prosodic reduction and the full redundancy of the verbally conveyed message. Given visibility, German speakers continued to use intensity as a prosodic marker to differentiate between the two accent types, even though intensity was overall reduced under visibility conditions. This finding again supports an integrated model of bimodal language production, where information transmitted visually (via a game move) is produced in congruence with information transmitted verbally.

Overall, we find support both for the effort-minimization view of speech production, leading to a gradient reduction in prosodic expression, and an integrated view of bimodal speech production, sustaining a categorical prosodic distinction across visibility conditions. However, it remains unclear to what extent our findings are caused by the presence or absence of visibility of the interlocutor, or the visibility of the game moves, as these two factors were confounded in our study. For now, we assume that the "prosodic reduction" effects we found under visibility are caused by the fact that the visual channel allows for an interpretation of the interlocutor's articulatory movements and visual prosody, thus aiding intelligibility and rendering more precise articulations unnecessary. We furthermore assume that the visual access to the game-related information did not affect production effort, as the fine-grained prosodic distinctions between different types of game moves were upheld under visibility. However, these assumptions needs further investigation. We therefore plan to collect further recordings involving visibility between interlocutors' faces, but without access to their game moves.

5. Acknowledgments

We would like to thank our participants who invested valuable lifetime to play a series of rather stupid games in the name of research, to the audience at the CRC colloquium at Saarland University for their valuable feedback, and to Denis Arnold who originally pointed out [6]'s study to us.

6. 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] 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.

[7] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility – head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, 2004.

[8] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 746–748, 1976.

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

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

[11] É. Lombard, "Le signe de lélévation de la voix," *Annales des Maladies de l'Oreille et du Larynx*, vol. XXXVII, no. 2, pp. 101–109, 1911.

[12] S. Kelly, A. Özyürek, and E. Maris, "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension," *Psychological Science*, vol. 21, pp. 260–267, 2010.

[13] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology. Journal of the Association for Laboratory Phonology*, vol. 3, pp. 71–89, 2012.

[14] S. Jannedy and N. Mendoza-Denton, "Structuring information through gesture and intonation," *Interdisciplinary Studies on Information Structure*, vol. 3, pp. 199–244, 2005.

[15] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and Speech in Interaction: An Overview," *Speech Communication*, vol. 57, no. Special Iss., pp. 209–232, 2014.

[16] P. Boersma and D. Weenink. (2016) Praat: doing phonetics by computer, version 6.0.22. [Online]. Available: http://www.praat.org

[17] 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/

[18] J. Peters, J. Hanssen, and C. Gussenhoven, "The phonetic realization of focus in West Frisian, Low Saxon, High German, and three varieties of Dutch," *Journal of Phonetics*, vol. 46, pp. 185–209, 2014.

[19] 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.