DOES INFORMATION-STRUCTURAL ACOUSTIC PROSODY CHANGE UNDER DIFFERENT VISIBILITY CONDITIONS?

Petra Wagner, Nataliya Bryhadyr, Marin Schröer, Bogdan Ludusan

Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies
Bielefeld University
petra.wagner@uni-bielefeld.de

ABSTRACT

It is well-known that the effort invested in prosodic expression can be adjusted to the information structure in a message, but also to the characteristics of the transmission channel. To investigate whether visibly accessible cues to information structure or facial prosodic expression have a differentiated impact on acoustic prosody, we modified the visibility conditions in a spontaneous dyadic interaction task, i.e. a verbalized version of TicTacToe. The main hypothesis was that visibly accessible cues should lead to a decrease in prosodic effort. While we found that - as expected - information structure is expressed throughout a number of acoustic-prosodic cues, visible accessibility to context information makes accents shorter, while accessibility to an interlocutor's facial expression slightly increases the mean f0 of an accent.

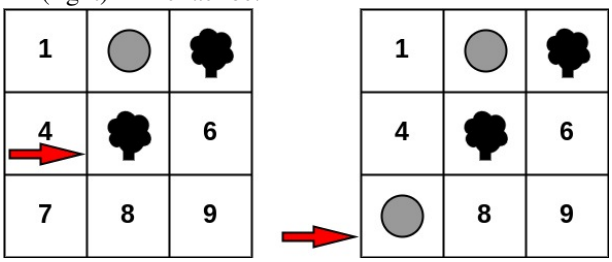
Keywords: prosody, dialogue, information structure, visibility, speech economy.

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 [10, 15, 2, 12].

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 [11]. Watson et al. [14] investigated whether different types of information-structure trigger different types of prosodic prominence. They operationalized the difference between *relevance accents* (roughly corresponding to "focus" in information theory) and *unpredictability accents* (roughly corresponding to "novelty" in information theory)) by measuring verbalized game moves in games 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 relevant, as they typically prevent the interlocutor from winning, or may constitute winning moves (cf. Figure 1). For American English, [14] 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. A replication of this study for another Germanic language, namely German, constitutes the first goal of this paper. We expect a similar accentual differentiation.

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



Most studies (including [14]) investigate the influence of information structure in ways neutralizing the influence of contextual factors that may affect information transfer, e.g. information that is accessible via a visual channel. However, in authentic communicative settings, interlocutors often can see their interlocutor's facial expression and have visible access to the information that is verbally expressed (e.g. in a TV weather report, speech is accompanied by a visual illustration). The interplay of visually and verbally accessible information in prosodic expression constitutes the second goal of our paper. Optimization models of speech communication would predict a decrease in overall prosodic effort if interlocutors have visible access to information that is verbally expressed [5, 7], but there are

several candidates that may trigger such an effect. When interlocutors see each other in the TicTacToe setting, they can see both the facial expression and the manually played game moves. Access to the manual game moves makes the verbal message fully redundant and might therefore lead to its pronunciation with reduced effort. However, visibility of facial movements have independently shown to enhance a message's intelligibility [9, 8, 4] and to convey prosodic prominence [4]. Therefore, a reduction in prosodic expression under mutual visibility could also result from exploiting a cumulative effect of visually and verbally conveyed prosodic prominence.

We therefore hypothesize that (1) if the visibly accessible information makes the verbal message redundant (which is the case if the manual moves to a game target field are visible), acoustic prosodic effort should be reduced. We further hypothesize that (2) if facial movements are visible (but not manual ones), these may provide cues to prosodic structure and thus cue a lack of vocal effort likewise. However, given the strong findings with respect to congruency across the visual and verbal modality [6, 3, 13], and given the fact that facially conveyed prosody does not render the verbal message redundant, we predict that the effect of facial visibility is less strong than the effect of visibility to relevant context information. Our subsequent study sets out to investigate our hypotheses.

2. METHODS

2.1. Recording Setup

The recordings were carried out at the faculty's recording studio using Sennheiser neckband microphones. We recorded 40 participants (native speakers of the Northern German Standard Variety) forming 20 dyadic pairs. Interlocutors within dyads were of equal social status, typically friends, and of same or mixed gender. Recordings of one speaker were excluded from further analyses due to technical problems resulting in a poor recording quality.

Each player received a set of cut outs in the form of blue or red felt squares to mark their moves on a shared vertical grid (cf. Figure 2). 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 and is realized sentence-finally in the vast majority of cases, e.g.

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

The verbalizations of game board targets ("1-9") were later analyzed with respect to their prosodic realization. However, speakers were not instructed to use a particular sentence structure or use specific words to refer to their targets.

2.2. Visibility Conditions

In each dyad, participants performed 4 games of TicTacToe in each of 4 different recording setups. The game board was placed horizontally between the speakers . In each game, the initial move was preset (randomly) by the experimenter. The order of game initializations rotated and the order of recording setups was shifted for each dyad. The 4 different recording conditions are specified in the following:

1. Manual and facial visibility: transparent game board, full view of interlocutor's head and facial expression.
2. Manual visibility, no facial visibility: transparent game board, but obstructed view of interlocutor's head and facial expression (with a light curtain)
3. Facial visibility, no manual visibility: non-transparent game board, but full view of interlocutor's head and facial expression
4. Neither manual nor facial visibility: non-transparent game board, obstructed view of partner's head and facial expression (with a light curtain)

Figure 2: The recording setup under the full visibility condition.

2.3. Information Structure Conditions

Largely following [14], we used the TicTacToe Setting to disentangle two aspects of information structure, namely *predictability* and *relevance*. As the initial moves were predefined by the experimenter, openly told to both participants and then simply repeated by the participant, they were defined as fully predictable ('8'). The second move was annotated as least predictable ('1'), as the participants still have many options to choose from on the game board. The following moves were annotated with increasing predictability ('2—8') in course of the game. For statistical analysis, predictability was recoded in a binary fashion, with predictability $> 4 =$ 'predictable', the remaining moves as 'unpredictable'. Relevance was operationalized in a binary fashion, with moves that prevent or constitute a winning move being annotated as 'relevant', others as 'irrelevant' (for the outcome of the game). As game decisive, relevant moves tend to come later in the game, and typically are predictable, the presence of either feature predicts the absence of the other, allowing for a contrastive analysis of both. For an illustration of *predictability* and *relevance*, cf. Figure 1.

2.4. Annotations and Analyses

In each dyad, the verbalizations of the game move targets (numbers 1-9) as well as the corresponding move's relevance and predictability were annotated manually using Praat [1]. In the vast majority of cases, these targets corresponded to the final word of an utterance, coinciding with a (nuclear) accent. We restricted our analyses to the phrase final, accented verbalizations of target moves. Using a Praat script, we then carried out a number of acoustic analyses of these target move verbalizations (duration (ms), mean f0 (st rel 1 Hz), f0 range (st), RMS intensity). These acoustic features served as dependent variables in the subsequent analyses. In a first step, it was then determined which participants were influenced by different visibility conditions by calculating a set of Linear Mixed Effect Models on all dependent variables, using the interaction of participant and both visibility conditions ('manual visibility', 'facial visibility') as fixed factors, and 'item' and 'dyad' as random factors. For the subset of participants who showed an interaction with visibility on any dependent variable, we then calculated a series of Linear Mixed Effects Models, using the above mentioned acoustic features as dependent variables, 'relevance', 'predictability', 'manual visibility' and 'facial visibility' as fixed factors, and 'item', 'participant' and 'dyad' as random factors. For a direct comparison of relevance and unpredictability accents, we coded a factor 'accent contrast', containing the labels 'important' and 'unpredictable', but excluding the cases where accents are neither 'unpredictable' nor 'relevant' (mostly initial moves), or both.

3. RESULTS

3.1. Influence of Visibility on Individual Participants

Out of the 39 remaining participant recordings, 37 showed an interaction effect between facial visibility or manual visibility on at least one of the examined acoustic parameters. Data from these 37 participants were entered into the analysis of potential effects of visibility and the prosodic expression of information structure.

3.2. Information Structure Effects

Both predictability and relevance have a significant influence on the duration of accented words, with relevant accents being significantly longer ($\beta = 15, SE = 5.36, t = 2.9$) and predictable accents being significantly shorter ($\beta = -32, SE = 5.37, t = -6.0$). When contrasting accents of unpredictability and accents of relevance, unpredictable accents are significantly longer than relevant ones ($\beta = 13.8, SE = 6.32, t = 2.18$, cf. Figure 3). These results are in line with previous research on American English [14]. Also, f0 range is influenced both by predictability and relevance, with relevant accents being produced with a higher f0 excursion ($\beta =$ 0.69, SE=0.31, t=2.2) and predictable accents with a more compressed one ($\beta = -0.9, SE = 0.31, t = -2.84$). However, a direct comparison of unpredictable and relevant accents fails to show a significant distinction: Both unpredictability and relevance increase f0 range in a similar fashion. This differs from results for American English, where unpredictable accents showed a larger f0 excursion. Globally, neither relevance nor predictability had an impact on RMS or Mean f0. Again, this finding differs from American English, where intensity was used to mark relevance accents rather than unpredictable accents.

3.3. Visibility Condition Effects

Visibility conditions had a significant influence on two acoustic parameters: If participants could see the interlocutor's hands, they produced their verbalized target moves significantly faster ($\beta = -10, SE = 5.14, t = -2.0$, cf. Figure 3). If partic-

Figure 3: Accent duration distribution depending on accent function (relevance vs. unpredictability accent) and absence (left) or presence (right) of manual game moves.

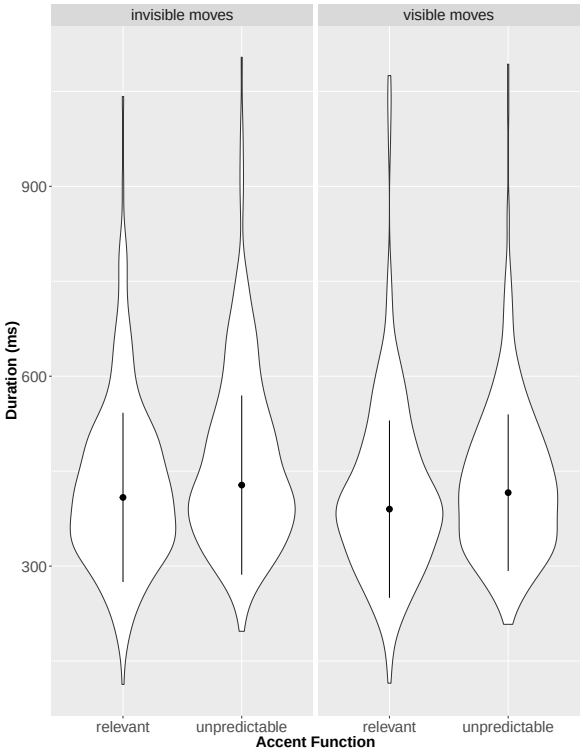
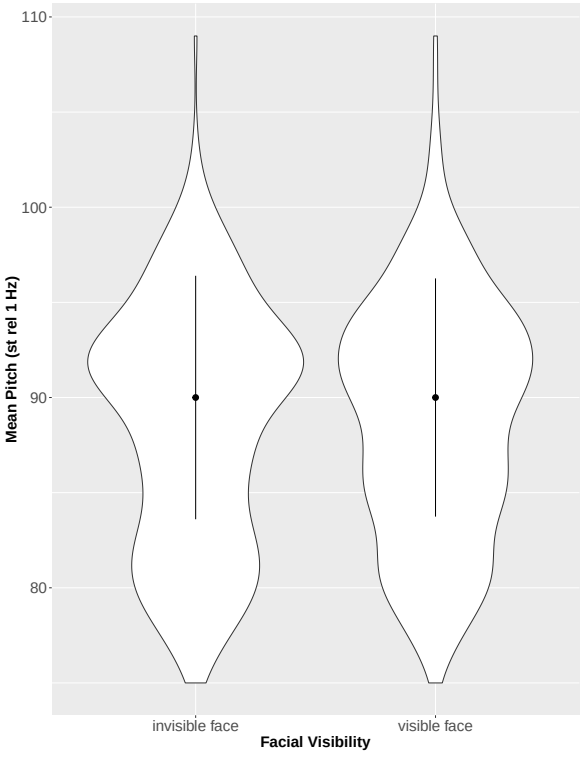


Figure 4: Distribution of mean f0 in accents both with and without access to facial visibility.



ipants could see each other's faces, they produced their target moves with a slightly higher average f0 ($\beta = 0.52, SE = 0.26, t = 2.0$, cf. Figure 4). RMS intensity was not affected by visibility, and neither was f0 range.

4. DISCUSSION

Accent functions are acoustically differentiated in German, but different from American English. For the latter, it was found that unpredictability accents are longer in duration and produced with a higher f0 range, but that relevance accents are louder. As in American English, German speakers produce unpredictable accents comparatively longer than relevant ones, but do not use f0 excursion to differentiate them further. Also different from American English, German speakers do not use intensity cues to signal accent function. As our analysis only looks at shallow acoustic features and does not take into account the fine detail of f0 movements or accent types, we refrain from concluding that predictability and relevance are not differentiated using f0 cues in German. A more fine-grained analysis of f0 movements is needed in the future.

With respect to the interplay of different visibility conditions and accent realization, we had hypothesized a comparatively stronger effect of manual visibility on the prosodic expression of accents. Indeed, this was supported by a decreased duration of accented words given access to manual visibility. If interlocutors have visible access to the information that is verbally transmitted, they invest less (durational) effort into their accent realization. Interestingly, no other prosodic features of prominence expression were affected, and f0 range stayed similar. Our findings thus provide some, but not full, support for speech optimization theories. Our second hypothesis was that visible access to the interlocutor's facial expression would also reduce prosodic effort, but less so. Here, we only found a very small positive impact of facial visibility on Mean f0, which could be interpreted as an increase in overall prosodic "engagement", possibly modulated by a strengthened interlocutor's co-presence. Our hypothesized reduced prosodic effort under facial visibility conditions was not corroborated by the results.

Maybe the most surprising result was the lack of influence of different visibility conditions on RMS intensity. We interpret this in such a way that speakers react with intensity modulations mostly to noise in the transmission channel, in a classic Lombard-like fashion. Lacking visibility of an interlocutor does not automatically create Lombard speech.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2019. Praat: doing phonetics by computer [computer program]. version 6.0.49.

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

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

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

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

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

[7] Lombard, É. 1911. Le signe de l'élévation de la voix. *Annales des Maladies de l'Oreille et du Larynx* XXXVII(2), 101–109.

[8] McGurk, H., McDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264(746–748).

[9] Munhall, K., Jones, J. A., Callan, D. E., Kuratate, T., Vatikiotis-Bateson, E. 2004. Visual prosody and speech intelligibility – head movement improves auditory speech perception. *Psychological Science* 15(2).

[10] Pierrehumbert, J., Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P., Morgan, J., Pollack, M., (eds), *Intentions in Communication*. Cambridge MA: MIT Press 271–311.

[11] Riester, A., Baumann, S. 2013. Focus triggers and focus types from a corpus perspective. *Dialogue and Discourse* 4(2), 215–248.

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

[13] Wagner, P., Malisz, Z., Kopp, S. 2014. Gesture and Speech in Interaction: An Overview. *Speech Communication* 57(Special Iss.), 209–232.

[14] Watson, D., Arnold, J., Tanenhaus, M. K. 2008. Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition* 106(3), 1548–1557.

[15] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics* (27), 55– 105.