THE ROLE OF LISTENER FEEDBACK IN PROSODIC CUE PRODUCTION: AN INTERACTIVE TASK

Lahari Chatterjee[1], Kathleen Schneider[1], Isabell Wartenburger[1], Sandra Hanne[1], Andrea Hofmann[1], Outi Tuomainen[1]

[1]University of Potsdam, Cognitive Sciences, Dept. Linguistics
lahari.chatterjee@uni-potsdam.de, kathleen.schneider.1@uni-potsdam.de, isabell.wartenburger@uni-potsdam.de, sandra.hanne@uni-potsdam.de, andrea.hofmann@uni-potsdam.de, outi.tuomainen@uni-potsdam.de

ABSTRACT

The aim of the study is to investigate the situational (in)dependence of production of prosodic boundary cues in an interactive setting. A game-like task was designed in which we tested whether native German speakers, when asked to repeat, change their speaking style while communicating with a confederate listener. In the task, they produced coordinate name sequences in German, with and without grouping of constituents, and received feedback from the listener. If prosody is used for the listener, we expect that after the feedback requesting a repetition of an item, speakers would switch from "casual" to "clear" speaking style.

Preliminary results showed that they use prosodic boundary cues (f0 range, final lengthening and pause) to mark groupings of the constituents but cue productions stayed consistent irrespective of listener feedback. Our results speak in favour of prosody being produced for the speaker and not primarily for the listener, i.e. situationally independent.

Keywords: boundary cue production, situational (in)dependence, listener feedback, clear speech.

1. INTRODUCTION

In dyadic face-to-face conversations, the listener signals understanding/misunderstanding of the message to the speaker via verbal and nonverbal feedback cues (e.g., clarification requests and back channelling) [1,2]. As a result, whenever there is a communication breakdown, speakers often adapt their speech and change from a conversational ("casual") to a more hyper-articulated ("clear") speaking style to increase the intelligibility of their speech [3,4]. These acoustic-phonetic adaptations often include reductions in speaking rate, increases in pause frequency, pause duration and fundamental frequency of speech [4]. Thus, it has been suggested that speech produced in interaction is highly dynamic and dependent on the perceived needs of the listener (i.e., listener-oriented and therefore situationally dependent) [5,6].

However, not all aspects of the speech signal are shown to be adaptive and listener-oriented [7, 8]. For example, a recent study by Huttenlauch et al. [9], investigated the adaptivity of prosodic boundary cues that signal syntactic grouping of three constituents in coordinate name sequences (henceforth coordinates). Speakers were asked to produce coordinates in two conditions (grouping vs no grouping) (see 1 and 2 below) and in five varying contexts corresponding to different virtual listeners introduced via a video recording (Young Adult, Child, Elderly, Non-native, Young Adult in Noise). It was assumed that if prosodic boundary cue production is listener-oriented, speakers will vary their productions between these different contexts.

(1) (Name1 und Name2) und Name3 (grouping)
(2) Name1 und Name2 und Name3 (no grouping)

The results showed, first, smaller f0 range and a decrease in final lengthening and pause durations of Name1; second, an increase in f0 range and final lengthening of Name2 as well as longer pause durations after Name2 in the grouping condition (1), relative to the no grouping condition (2). These findings are in line with the syntax-prosody model proposed by Kentner and Féry [10]. Yet, in terms of the five different contexts, speakers showed only limited adaptations to the different interlocutors. The authors concluded that prosody is not produced with the listener's needs in mind, but it is speaker-oriented and situationally independent [7,8]. However, the study by Huttenlauch et al. [9] relied on virtual listeners who provided no feedback to the speaker. Therefore, it is possible that these "passive" listeners were not real enough to elicit listener-oriented speech adaptations. To account for this, the present study investigates situational independence of prosodic boundary cue production in an interactive game-like setting with a real interlocutor (a confederate listener) who provides real-time feedback on the communicative success (understood/ misunderstood/ repeat request).

We hypothesize that, if prosodic boundary cue production is listener-oriented and situationally dependent, speakers will adapt their speech to the

confederate's feedback and produce casual speech when the confederate listener understands them correctly. However, if the confederate listener requests for a repetition, they would change their speaking style to clear speech. This is expected to result in enhanced prosodic boundary cue productions (e.g., increase in f0 range and final lengthening of constituents and longer pause durations). If, however, prosodic boundary cue production is situationally independent, the speaker would not be influenced by the confederate listener's feedback and prosodic cue productions would remain consistent regardless of whether the listener asked them to repeat.

2. METHOD

2.1. Participants

So far, we recorded 5 adult native speakers of German (male and female, mean age = 26.6. years, *SD* = 5.59; we are aiming for a final sample of 30 speakers). All participants reported normal hearing and no history of language impairments. Informed consent was obtained from the participants and they were remunerated or received course credits for participation. The study was approved by the ethics committee of the University of Potsdam.

2.2. Stimuli

Experimental items were adapted from [9] and [11]. As we needed a bigger number of foil items to test the two speech styles (casual/clear) relative to the listener feedback type (understood/misunderstood/repeat), we added more names in addition to those used in [9, 11] to construct the name sequences. Experimental items included seven coordinate name sequences consisting of three names coordinated by '*und'* (*and* in English). Each name (n=15) used in the sequences was disyllabic and trochaic. Seven of the total of fifteen names, ending in the high front vowel /i/ to reduce glottalisation [9] (Gabi, Manni, Mimmi, Moni, Nelli, Leni, Lilli), were used as Name1 and Name2 for the experimental items. Four names (Manu, Nina, Lola, Lisa) appeared as Name3 in the experimental and filler items, while four additional names (Dora, Lotte, Mira, Sara) served only as filler items.

Each name sequence appeared in two grouping conditions, grouping and no grouping (see examples 3-4).

2.3. Procedure

In the course of the experiment, the speaker sat inside a sound-attenuated audio booth, while the confederate listener (member of the research team) was seated outside the booth. The interlocutors could see and hear each other via headphones/microphones. In order to make the listener's miscomprehension more realistic, the speaker was informed that the confederate listener heard background noise (20-talker BABBLE noise) during the task.

The experiment began with a practice phase consisting of seven example trials where the participants were familiarised with the names at first and then with the two grouping conditions (see examples in 3 and 4). Thereupon, the test phase consisted of 4 blocks of 120 trials with 3 self-paced breaks every 30th trial. During the breaks, in order to make it more realistic, the interlocutors were given a chance to interact with each other and discuss their performance.

In the introduction phase, the speaker was informed that the listener would do an auditory word-picture-matching task. In the test phase, the task began with a question "*Wer kommt?*" (*Who is coming*?), followed by a fixation cross for 1000 ms. Next, the speaker was shown three pictures, each consisting of a visual presentation of three name sequences as well as the corresponding name sequence (see 3-5) written underneath the picture. There was one target picture with the corresponding name sequence written underneath (3) and two other pictures and sequences displaying a different grouping condition with the same names (4), or the same grouping condition with different names (5), respectively. The speaker was then prompted to produce the highlighted target sequence with the intended grouping (either with or without grouping; see examples 3-4).

(3) *(Moni und Lilli) und Lisa*
(4) *Moni und Lilli und Lisa*
(5) *(Lola und Manu) und Leni*

There were three forms of listener feedback: i) the confederate listener understood the speaker correctly and the speaker received feedback on their screen in which a green tick appeared above the target item and the next trial followed; ii) the confederate listener misunderstood the speaker and a red cross appeared above the misunderstood item and the next trial followed; iii) the confederate listener requested a repetition of the item in which case an orange question mark appeared on the screen. The analyses focus on the coordinate sequences produced in the feedback condition iii, where we expect the speaker to change their speaking style from casual to clear when they repeat the item. The task took approximately thirty minutes in total. Both in the practice and the test phase the confederate listener responses were automated to ensure that on approximately 11% of randomly introduced trials, the

confederate listener requests for a repetition. A feedback questionnaire at the end confirmed that the speakers were not aware of this manipulation.

2.4. Annotation and data processing

Out of the total of 140 productions (7 name sequences * 2 grouping conditions (grouping and no grouping) * 2 speaking style (casual and clear) * 5 speakers), 107 productions, recorded at a sampling rate of 44100 Hz (16 bit), were manually annotated using Praat [12]. Due to technical issues during the recording, 33 productions were discarded. Three acoustic cues were measured on Name1 and Name2: f0 range, final lengthening of the last vowel, and pause duration; identical to the criteria employed by Huttenlauch et al. [9]. For f0 range, the range between f0 min and f0 max was measured in semitones formula used for calculation: $12*\log_2(f0_{max}/f0_{min})$ [9]. Further, f0 was time normalised and the resulting contours are plotted in Hz as seen in Figures 1-3. Final lengthening was calculated as the duration of the final vowel (i.e., the last segment) of Name1 and Name2 relative to the total duration of Name1 and Name2 (in %), respectively. Lastly, pause duration was measured as the duration of the pause following Name2 relative to the total duration of the utterance (in %). Preliminary statistical analysis was run on the percentage scores.

First, we report the results for the grouping conditions (grouping vs no grouping). Second, we report the results for the two speaking styles (casual vs clear) in response to the listener feedback. Our dependent variables were f0 range, final lengthening and pause and our independent variables are the two grouping conditions (grouping vs no grouping) and speaking style (casual vs clear). Because in this paper we report preliminary data from only 5 speakers, the statistical analyses focus on the main effects of the grouping condition and speaking style (and not their interaction), for each dependent variable separately (a non-parametric Mann Whitney U test). Results from the linear mixed effect modelling for the full dataset and all comparisons will be reported in the poster.

3. RESULTS

Preliminary results showed a larger f0 range on Name1 in the no grouping condition *(M=7.06, SD= 3.55)* compared to the grouping condition *(M=4.80, SD= 4.85)* (*p<0.001*) (see Figure 1). With regards to speaking style, there was no significant difference (*p*=0.067) (see Figures 2 and 3).

In addition, f0 range on Name2 in the grouping condition was larger (*M=11.51, SD=5.27*) than the f0 range on Name2 (*M=6.77, SD=5.20*) in the no grouping condition (*p<0.001*) (see Figure 1). However, there was no significant difference with regards to the speaking style (*p=0.189*) (see Figures 2 and 3).

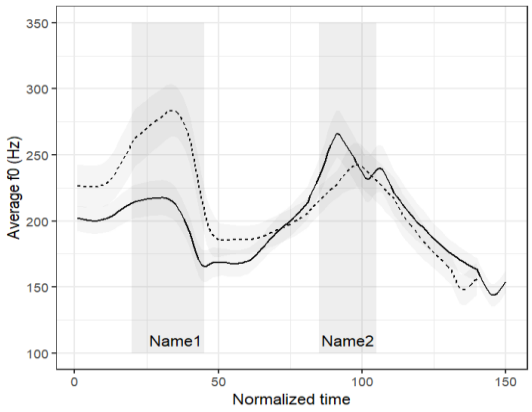

Figure 1: Time normalised f0 (in Hz) contours for the two grouping conditions. Solid line indicates the grouping condition and dashed line the no grouping condition.

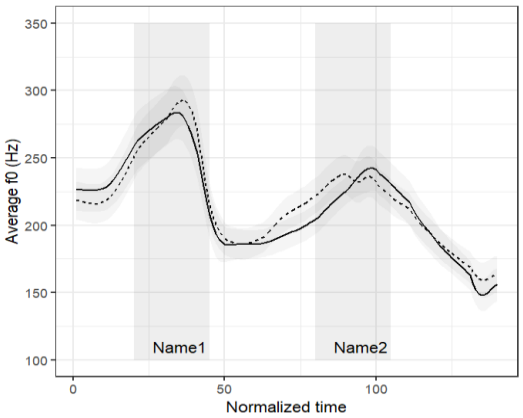

Figure 2: Time normalised f0 (in Hz) contours for the two speaking styles in the no grouping condition. Solid line indicates the casual speaking style and dashed line the clear speaking style.

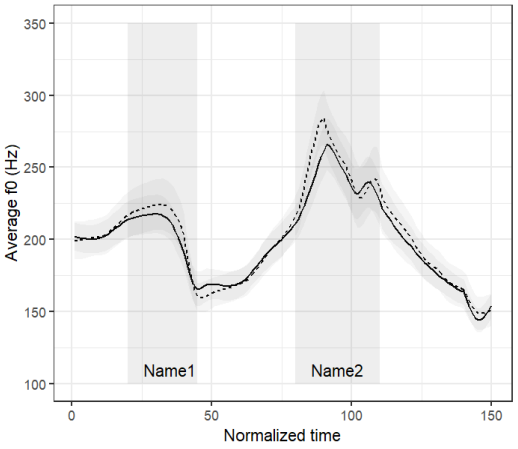

Figure 3: Time normalised f0 (in Hz) contours for the two speaking styles in the grouping condition. Solid line indicates the casual speaking style and dashed line the clear speaking style.

Moreover, the duration of the final segment of Name1 in the grouping condition shortens (*M=31.15, SD=4.85)* as compared to the final segment duration on Name1 (*M=34.90, SD=4.57*) in the no grouping condition (*p<0.001*). But final segment lengthening of Name1 showed no statistically significant difference between speaking styles (*p=0.062*).

Further, the duration of the final segment of Name2 in the grouping condition lengthens (*M=42.20, SD= 5.69*), as compared to final segment duration on Name2 (*M=34.08, SD=5.46*) in the no grouping condition, with *p<0.001*. Relative to the speaking style, we observed a statistically significant effect (*p<0.05)*. The duration of the final lengthening of Name2 in clear speaking style (*M= 39.44, SD=6.60*) is longer than the casual style (*M=36.46, SD=6.91).*

Lastly, the pause after Name2 was longer (*M=16.49, SD=10.68*) in the grouping condition as compared to the no grouping condition (*M=0.20, SD=0.85*) showing a significant difference *(p<0.001).* However, no statistically significant difference was found for the speaking style (*p=0.719*).

In sum, results showed statistically significant effects of the grouping condition on the f0 range on Name1 and Name2, final lengthening on Name1 and Name2, relative to the no grouping condition. In addition, there was a significant effect of the grouping condition on the pause duration on Name2. However, we found limited changes in regards to speakers modifying their speaking style as a response to the listener feedback. However, these results need to be validated with more data.

4. DISCUSSION

The aim of the study was to investigate the situational (in)dependence of prosodic boundary cues in an interactive setting. We investigated how speakers modify their speech from casual to clear speaking style after they were requested to repeat an item by the confederate listener. Preliminary results with five speakers showed that the three cue measurements (f0 range, final lengthening and pause) clearly mark the difference between the two grouping conditions indicating a close and stable syntax-prosody link. These results replicate those reported by Huttenlauch et al. [9], Kentner and Féry [10], Petrone et al. [13].

However, with regards to the speaking style, our results showed that there are no changes, or only limited changes in some of the cues (e.g., Name2 final lengthening), in the three acoustic cues between casual and clear speaking styles, thus, partially replicating the results by Huttenlauch et al. [9] obtained with "virtual" interlocutors, now with a "real" interlocutor. It is worth noting that here we reported data from only five speakers and it is, therefore, possible that with a larger sample of speakers we will see listener-oriented adaptations also in (at least some) prosodic boundary cues. As a next step, for the casual-to-clear speaking style modifications, we will run further exploratory analyses for cues other than f0 range, final lengthening and pause, and places other than Name1 and Name2.

As our preliminary data indicates that speakers do not modify their prosodic cue production in response to the listener feedback (apart from one cue), our data support the assumption that prosodic boundary production is independent of the listener. In the context of listener-oriented accounts [5], these results can be somewhat surprising. For example, it is well known that speakers increase their pitch range when speaking to infants and smaller children in contrast to when they are speaking to another adult [14]. However, it is possible that at the prosody-syntax interface, in specific contexts, prosody might serve different functions and could be unaffected by conversational situations as suggested by previous studies [10,15]. Alternatively, it could also be that the speakers were already maximally marking the prosodic boundaries (i.e, at ceiling) in the casual style. Thus, the fact that prosodic cue productions are independent of the interlocutor or the communicative situation, speaks in favour of prosody being situationally independent and mainly produced for the speakers themselves. However, so far, we have tested only a small and preliminary sample, and more data is needed to validate these results.

5. REFERENCES

[1] Krauss, R. M., Garlock, C. M., Bricker, P. D., McMahon, L. E. 1977. The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology* 35(7), 523.

[2] Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R. 2007. Fluid Semantic Back-Channel Feedback in Dialogue: Challenges and Progress. In: Pelachaud, C., Martin, JC., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds) *International Workshop on Intelligent Virtual Agents*. Springer, 154-160.

[3] Buz, E., Tanenhaus, MK., Jaeger, TF. 2016. Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89, 68-86.

[4] Tuomainen, O. T., & Hazan, V. 2018. Investigating Clear Speech Adaptations in Spontaneous Speech Produced in Communicative Settings. In: M. Gósy, T. E. Gráczi (Eds.), *Challenges in analysis and processing of*

spontaneous speech. Budapest, Hungary: MTA Nyelvtudományi Intézet, Budapest.
[5] Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In: W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* Dordrecht. Springer, 403–439.
[6] Smiljanić, R., and Bradlow, A. R. 2009. Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass* 3, 236–264.
[7] Clifton, C., Carlson, K., Frazier, L. 2002. Informative prosodic boundaries. *Language and Speech* 45, 87–114.
[8] Speer, S. R., Warren, P., & Schafer, A. J. 2011. Situationally independent prosodic phrasing. *Laboratory Phonology* 2(1), 35–98.
[9] Huttenlauch, C., et al. 2021 Production of prosodic cues in coordinate name sequences addressing varying interlocutors. Laboratory Phonology: *Journal of the Association for Laboratory Phonology* 12(1), 1–31.
[10] Kentner, G., & Féry, C. 2013. A new approach to prosodic grouping. *The Linguistic Review* 30(2), 277–311
[11] De Beer, C., Regenbrecht, F., Huttenlauch, C., Wartenburger, I., Obrig, H., Hanne, S. 2021. Kommunikative Beeinträchtigungen nach rechtshemisphärischer Hirnläsion: Produktion und Perzeption von Prosodie. In Jaecks, P., Hussmann, K., Mantelli, A., Monaco, E., Ruflin, D. (eds). *Aphasie und verwandte Gebiete | Aphasie et domaines associés N°1/2021*.
[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.01.
[13]Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., & Höhle, B. (2017). Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics* 61, 71–92.
[14] Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. 2002. Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior & Development, 24*(4), 372–392.
[15] Kraljic, T., & Brennan, S. E. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50(2), 194–231.