Perception of F0 movements towards potential turn boundaries in German and Swedish conversation: background and methods for an eye-tracking study

Martina Rossi[1], *Kathrin Feindt*[1], *Margaret Zellers*[1]
[1] *Kiel University, Germany*
mrossi@isfas.uni-kiel.de, kfeindt@isfas.uni-kiel.de, mzellers@isfas.uni-kiel.de

Abstract

Understanding the turn-taking system in conversation entails not only knowledge about the linguistic-structural and phonetic before Potential Turn Boundaries (PTBs), but crucially, the precise location of the transition space as well. To investigate the time domain and the phonological domain, we compare production and perception of turn ends in two related languages: German and Swedish. For the first part, we extracted pitch values at seven time points before PTBs from spontaneous speech produced in two-party conversations. The aim was to investigate the possible presence of specific patterns of variations that lead to either speaker change, floor keeping or backchannels. As no such patterns have emerged, for the second part, eye-tracking will be used to investigate the exact time point at which the ending of a turn can be projected by a listener and which acoustic signals are important for this prediction.

Introduction

Face-to-face conversation is a fundamental part of human social behavior. Verbal interactions between two (or more) interlocutors are generally characterized by one speaker talking at a time and the interlocutor intervening at the appropriate moment, typically avoiding marked silent gaps or long stretches of overlapped speech (Sacks et al., 1974).

The achievement of smooth turn exchanges entails that listeners are ready to launch their turn as soon as the speaker has finished talking, suggesting that they should be able to predict the current speakers' intentions of either holding or ceding the floor early within the current turn in order to start preparing their next conversational move (Levinson & Torreira, 2015). Comparing the observation that the most frequent silent gap has an average duration of around 200 ms (Stivers et al., 2009; Heldner & Edlund, 2010) with evidence from language production studies supports the hypothesis of early planning in turn-taking: in fact, the minimum time for the encoding of a single word is around 600 ms (Indefrey & Levelt, 2004; Schnur et al., 2006; Indefrey, 2011), while at least 1500 ms are needed for the planning of a short clause (Griffin & Bock, 2000). Predictions by listeners are made on the basis of their perception of several cues (linguistic, phonetic, gestural), produced by the current speaker, which signal the approach of a Potential Turn Boundaries (PTB Zellers, 2017), i.e. a location where speaker change becomes possible, though not obligatory (see e.g. Transition Relevance Places; Sacks et al., 1974). How early cues' variation is related to turn-taking starts, and thus what the precise location of the transition space is, is still a matter of debate. While De Vos, Torreira & Levinson (2016) argue for the last 500 milliseconds for the projection of a turn end, Zellers (2017) found relevant phonetic signals up to 1 second before a PTB. Among prosodic-phonetic turn-taking cues, F0 movements towards potential boundaries have been observed to play an important role in signaling to the listener the intentions of the current speaker (Couper-Kuhlen & Selting, 1996; Hjalmarsson, 2011; Bögels & Torreira, 2015; Heldner & Włodarczak, 2015, *inter alia*). However, the extent of F0 contribution to the turn-taking system seems to be constrained by the prosodic phonology of the language object of study (Zellers, 2014; Zellers et al., 2019).

The present research has three main objectives: (1) analyze the patterns of variation of F0 towards PTBs and its contribution to turn-taking in spontaneous conversation in German and Swedish, i.e. two languages which are related but differ in their prosodic phonology; (2) investigate the extent of the transition space by monitoring F0 at several test locations approaching a PTB; (3) test the production results through an eye-tracking experiment involving native speakers of the two languages analyzed. This paper reports the first results of the production study carried out on German and Swedish interactions, and describes the methods, the design and the preliminary hypotheses for the eye-tracking study.

Materials and Methods

For the production study we analyzed two-party spontaneous conversations in German and in Swedish. As previously discussed (see Introduction), German and Swedish have been selected to investigate F0 movements as a turn-taking cue cross-linguistically since their prosodic phonologies are typologically different. On one hand, German is an intonation language, where pitch accents mark prominent syllables as well as carry information about the discourse structures, and boundary tones generally mark the end of an intonational phrase; on the other hand, Swedish is a pitch accent language (Gårding, 1989), where two different lexical pitch accents contrast many word pairs, and focus is signaled through an additional H tone after the pitch accent of the focused item. Moreover, Swedish intonation phrases are marked by boundary tones, similarly to German, even though in Swedish they tend to be an L% in most cases (House, 2004; 2005). These differences between the two languages may have some interesting influence on the variation of F0 as a turn-taking cue in spontaneous conversation.

The German dialogues investigated come from the Lindenstraße task of Kiel Corpus of Spoken German (Kohler et al., 2018), while the Swedish data are taken from the Spontal Corpus (Edlund et al. 2010). For both corpora, the two speakers that took part in the experimental session were recorded through two separate microphones, which allows a phonetic analysis of the speech signal even when the two interlocutors talk simultaneously. We have annotated and analyzed 2 conversations for German, with 4 different speakers, and 2 conversations for Swedish, with 4 different speakers, 10 minutes for each conversation. The current sample of speakers was not balanced for gender: in the 2 Swedish conversations analyzed we have 3 male speakers and 1 female speaker, while in the 2 German dialogues we have 4 female speakers.

Roughly 500 PTBs were identified and manually annotated using Praat (Boersma & Weenink, 2022) applying a set of labels (Feindt, Rossi &, Zellers, 2021; Rossi, Feindt & Zellers, forthc.) describing what came after the potential boundary. First of all, a label was assigned to describe the completion of the utterance in context on a pragmatic/syntactic level ("yes" for complete and "no" for incomplete utterances); a label was assigned to describe the sentence type ("d" for declarative utterances, "q" for questions, "t" for tag questions). Then, a label was assigned to describe which sequential structure occurred after the PTB:

"c", for change, is assigned to those cases where the other interlocutor took the floor after the PTB;

"k", for keep, is assigned to those cases where the current speaker held the floor after the PTB;

"b", for backchannel, is assigned to those cases where the other interlocutor produced a minimal, non interrupting response after the PTB.

Finally, a label referred to the transition type was aimed at describing the way in which the change, keep, or backchannel production took place:

"g", for gap, identified a transition accompanied by a silent gap longer than 120 ms;

"o", for overlap, identified a transition accompanied by a speech overlap longer than 120 ms;

"n" , for no-gap-no-overlap, identified a smooth transition, with possible silent gaps or speech overlaps with a duration inferior to 120 ms (Heldner, 2011).

After the annotation, we extracted F0 values using Praat's setting for semitones above 1 Hz. Data points were extracted using a script at several locations within the current turn: (i) at the PTB, (ii) 200 ms, (iii) 400 ms, (iv) 500 ms, (v) 700 ms, (vi) 900 ms, (vii) 1 s before the PTB. To avoid the influence of physiological factors on F0 measurements, we normalized the data with the speakers' individual baseline, calculated following the procedure used by Zellers & Schweitzer (2017).

Results

Most importantly for our research question, there were no correlations found between certain F0 measurements and a specific type of sequential structure. Due to a high degree of inter-speaker variability, no clear pattern of variation for F0 contour shape emerged before speaker changes, turn holds or backchannels. Similarly, we were not able to highlight any specific patterns leading up to a gap, an overlap or a no-gap-no-overlap in the transition between turns. Observing the interaction between the controlled variables, however, gave us some insight into the high degree of variability. For example, it appears from our data that, for speaker change cases, the syntactic completeness of the utterance influences the degree of variability of the F0 data points: in fact, in both languages, when the (upcoming) completeness is signaled by the syntax, we observed that, before no-gap-no-overlaps and overlaps, the degree of variability was a lot higher than when the utterance was not syntactically complete. This suggests that speakers are free to vary more with F0 movements when the completeness of their utterance is signaled by other communicative means, e.g. syntax (Selting, 1996).

Moreover, in speaker change cases, we noticed how the variability degree in F0 was smaller when the PTB was followed by a gap. In such cases, German speakers ended their utterance at around 13 st while Swedish speakers ended at around 2.5 st above the speakers' baseline, respectively higher than the preceding data points and lower than the preceding data points. Since these PTBs were followed by a gap, we can hypothesize that the interlocutors interpreted the final rise for German and the final fall for Swedish as a turn hold cue, so they did not intervene, even if the current speaker had the intention of ceding the floor. Comparing the data from the keep cases, we see that, similarly, German speakers' F0 at the PTB is around 14 st, while it is at 3 st for Swedish speakers. A final higher F0 might thus be a turn hold cue for German subjects, while the same intention might be signaled by a final lower F0 for Swedish subjects.

Even if specific patterns did not emerge in German or Swedish, we were able to make some interesting observations through a cross-linguistic comparison. We observe that German speakers end their turns higher compared to Swedish speakers. For example in declarative utterances, the Swedish mean values for the measuring point directly at the turn boundary lies 3 st above the baseline, while German speakers end higher, at 11 st on average, when followed by a speaker change case (Feindt, Rossi & Zellers, 2021). Additionally, there is greater variation in the German speaker's F0 span of data points compared to Swedish speakers, who end closer to their baseline and exhibit a much more coherent intonational pattern towards PTBs. Thus, again, in declaratives before a speaker change, the standard deviation (SD) in German is 5.34 st, while in Swedish the SD is 3.38 st (Feindt, Rossi & Zellers, 2021). Moreover, we find evidence for possible accommodation tendencies in F0 values between conversational partners in German, though not in Swedish. In fact, speakers interacting with each other in the two German conversations analyzed seemed to use a very similar range of F0 values in the test locations closer to the PTB and at the boundary, in both speaker change and keep cases, which could possibly be evidence for a local entraining behavior in the use of F0 as a turn-taking cue (Rossi, Feindt & Zellers forthc.). For instance, for speaker change, the first speaker pair's average for the three last F0 data points (at the end of the utterance, at 200 ms and at 400 ms from the PTB) is around 15 st, while for the second speaker pair it is at around 5 st. Similarly, in the keep condition, participants in the first conversation show an average of 1 5st for the final data points, while the other pair's average is at around 5 st, both with some degree of variation. Accommodation tendencies in F0, among other phonetic cues, have already been observed in previous studies on conversational exchanges (e.g. Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014). Even if our observation is based on qualitative data from a small sample of speakers, we hypothesize the possible presence of an accommodating behavior in F0 values, that will have to be further tested with more specific entrainment measurements. However, we are able to exclude the influence of a physiological similarity thanks to the normalization of our data with the individual speaker's baselines. As previously mentioned, the same similarities in the distribution of F0 data points among speaker pairs were not observed in the Swedish interactions. This may suggest that Swedish subjects in this data sample are not using F0 to accommodate with each other, but it does not exclude the possibility that, if they are entraining with each other, they might be relying on other phonetic features.

Discussion and Further Steps

Generally, our results show that the Swedish pitch accent indeed has an influence on turn final F0 patterns. German speakers show a higher variability of F0 range across all test locations and conditions. Taken together with the observation that there might be some cases of entraining of F0 between German speakers – which was not observed for Swedish – we concluded that the Swedish lexical pitch accent system hinders using F0 as a resource for communicative purposes, as it is saturated with phonological information already. As German does not have a phonological lexical pitch accent, speakers have more freedom to use their entire F0 span even for purposes of accommodation, for example.

With regard to our research question addressing the relationship between the time and the phonological domain for signaling upcoming turn ends, we could not pinpoint specific time points in the last second of a turn at which pitch is used as a signaling cue. No clear intonation patterns emerged from our analysis that rather lead to speaker changes, floor keeps or backchannels. Likewise, the phonological configuration does not seem to influence whether a smooth transition, a gap or an overlap followed. As mentioned earlier, it is nevertheless assumed that a turn at talk entails such signaling cues, though, that enable the listener to make predictions about the upcoming turn end. Those cues and their timing will be investigated further in a subsequent perception study. For this, eye-tracking is used to investigate the exact time point at which the ending of a turn can be projected by a listener. We use a simplified visual world paradigm in which a participant follows an excerpt conversation via headphones. The participant acts as a passive third party, but is prevented from seeing the conversational partners, though, to exclude gestures and mimic as turn-taking cues. Instead, listeners are presented with humanlike avatars that stand for the respective speakers in the conversation. The main task is to predict which speaker will talk next when a turn comes to an end by clicking on the corresponding avatar. Tracking the clicks allows us to assess the offline processing as well. In fact, the amount of time between the gaze shift and the click shows us how quickly the decisions are made. A greater time span for some stimuli compared to others may indicate an ambiguous phonological gestalt of the formers, resulting in a delayed click. A further, more practical use of the clicks is to determine whether the subjects' prediction corresponds to the actual course of the conversation excerpts taken from the corpora, i.e. if speaker changes and keeps are actually recognised as such.

The stimuli for the experiment come from the previously annotated interactions used for the production study, and they have been balanced for the sequential structure that followed the PTB (a speaker change or a turn hold; backchannel cases are not included at this stage). The PTB label also represents the end of each individual stimulus, after which the participants in the eye-tracking experiment have to make their decision about which speaker will talk next. The stimuli will be manipulated in (i) the overall pitch in the last 500 ms (raised and lowered), (ii) the last pitch accent (raised and lowered), (iii) the loudness (increased and decreased), (iv) speech rate (sped up and slowed down). Thus, we can assess which version sounds more like the speaker will continue speaking, or which version leads the participants to project a speaker change. More importantly, through the eye-tracker, we can determine the point in time at which the projection was made by the participant, as that will be the time at which the gaze shifts from the current speaker to interlocutor-avatar anticipating this is where the next turn will come from. In order to determine whether certain linguistic signals are prioritized in different languages, this part of the project will be carried out with Swedish as well as German native speakers. That these groups behave differently in the production of turn ends has been shown in our previous investigation. Hence, it needs to be investigated whether the perception will be influenced by the native language as well. The eye-tracking method will discern whether it is actually a certain time point before the end of a PTB that listeners use to project the upcoming turn end or rather phonological cues. A simplified pilot study in Germany revealed that participants were indeed able to use signals in the current speech to guide their predictions. Testing only turn keep versus turn yield cases – without further manipulation – showed that the gaze shifted to the "listening" avatar prior to the turn end. Crucially, this was not a fixed time point for all stimuli but rather appeared to correlate with the pitch accents, as the gaze shifted to the other avatar briefly after those. Although these are first, tentative results, they are nevertheless promising outcomes to be supported by further research.

Conclusions

There are two different approaches to the study of turn-taking signals: the time domain and the phonological domain. In the first part of our study, we investigated whether there are specific time points in the last second of a turn at which F0 is used as a cue for turn-taking. Analyzing roughly 500 PTBs in two languages led to the conclusion that there are no specific F0 values or shapes related to specific time points signaling either the intentions of holding or ceding the floor. Important results were however obtained by comparing German to Swedish turn-taking. Swedish, as a pitch accented language, shows less variation in the F0 range toward a PTB than German. Because pitch does not serve a phonological function in German, it can have a communicative purpose. This is for example manifested when speakers adjust their pitch to each to show a degree of accomodation with the conversational partner. These cross-linguistic differences will also be considered in a perception study that will attempt to assess the importance of specific time points in relation to phonological cues in turn-taking. The processing of turns at talk by a silent third party will be assessed using eye-tracking. Analyzing the gaze of participants when hearing excerpts of conversations allows us to pinpoint the exact time at which a speaker change is expected to happen, in that the gaze will shift to the interlocutor in anticipation of his turn. Moreover, we will be able to judge the importance of certain phonological parameters by controlling pitch height, the height of the pitch accent, speech rate and loudness. Our work thus contributes greatly to the ongoing discussion about turn-taking signals by adding a further dimension: direct online processing of natural speech.

Acknowledgments

This research was supported by the project Sound Patterns and Linguistic Structures at the Transition Space in Conversation (ZE 1178/3-1), funded by the German Research Council (DFG). Many thanks to our student assistants Sophie Caliebe and Oke Matthiesen for their work on the annotations.

References

Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.13, retrieved 18 May 2022 from http://www.praat.org/

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52, 46-57.

Couper-Kuhlen, E., & Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on. *Prosody in conversation: Interactional studies*, 11.

de Vos, C., Torreira, F. J., & Levinson, S. C. (2016). Turn-timing in signed conversations: coordinating stroke-to-stroke turn boundaries. *Frontiers in Psychology*, 6: 628.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. In *LREC* (pp. 2992-2995).

Feindt, K., Rossi, M., & Zellers, M. (2021). The time course of pitch variation towards possible places of speaker transition in German and Swedish. Proceedings of the 1st International Conference on Tone and Intonation (TAI) 2021.

Gårding, E. (1989). Intonation in Swedish. Working papers/Lund University, Department of Linguistics and Phonetics, 35, 63-88.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274-279.

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1), 508-513.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555-568.

Heldner, M., & Włodarczak, M. (2015). Pitch slope and end point as turn-taking cues in Swedish. In 18th International Congress of Phonetic Sciences, Glasgow, Scotland, UK, August 10-14, 2015. University of Glasgow.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1), 23-35.

House, D. (2004). Final rises and Swedish question intonation. Proceedings of FONETIK 2004, 56-59.

House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human–machine dialogue. *Speech Communication*, 46(3-4), 268-283.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Frontiers in Psychology*, 2, 255.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2), 101-144.

Kohler, K. J., Peters, B., & Scheffers, M. (2018). The Kiel Corpus of spoken German: Read and spontaneous speech.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In INTERSPEECH.

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (pp. 5-12).

Rossi, M., Feindt, K., & Zellers, M. (forthc.). Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish. Proceedings of Speech Prosody 2022.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, 50, 696-735.

Schnur, T. T., Costa, A., & Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of psycholinguistic research*, 35(2), 189-213..

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* 6, 357-388. Journal of psycholinguistic research, 35(2), 189-213.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106(26), 10587-10592.

Zellers, M. (2014). Duration and pitch in perception of turn transition by Swedish and English listeners. In Proceedings of FONETIK (pp. 41-46).

Zellers, M. (2016). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech* 60(3): 454-478.

Zellers, M., & Schweitzer, A. (2017). An Investigation of Pitch Matching Across Adjacent Turns in a Corpus of Spontaneous German. In INTERSPEECH (pp. 2336-2340).

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019). Timing properties of hand gestures and their lexical counterparts at turn transition places. In Proceedings of FONETIK (pp. 119-124).