POINTS OF MAXIMUM GRAMMATICAL CONTROL – THE PROSODY OF A TURN-HOLDING PRACTICE

{Anneliese Kelterer, Saskia Wepner}*, Julian Linke, Barbara Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology
anneliese.kelterer@edu.uni-graz.at, saskia.wepner@student.tugraz.at, {linke, b.schuppler}@tugraz.at

ABSTRACT

In conversation, speakers may pause in the middle of a turn construction unit to hold their turn, at a point of *maximum grammatical control*. We call such turn holds 'incomplete holds'. Based on 837 inter-pausal units from a corpus of Austrian German spontaneous conversations, we extract 11 prosodic features (F0 and RMS, one durational feature and one categorical feature) to analyse the prosody of incomplete holds. We do so in two steps: First, we use a Random Forest classifier to find the most important features that contribute to distinguishing incomplete holds from other kinds of pre-pausal turn-holds as well as from turn-changes. Second, we use regression analysis to examine how these features contribute to the different turn-taking categories. We find that incomplete holds are characterised by continuing intonation, high intensity, low articulation rate and a flatter F0 than grammatically complete holds.

Keywords: Turn-taking, prosody, conversational speech, Austrian German, incomplete hold

1. INTRODUCTION

Speakers often produce pauses at the end of turn construction units (TCU) [1]. However, speakers may also pause within a TCU to hold their turn, at a point of *maximum grammatical control* [2] (e.g., "I understand it, but ... no one else does"), called incomplete holds in this paper. This phenomenon has rarely been studied. One notable contribution is the *qualitative* investigation of turn-holding and turn-yielding practices in a North-Western variety of German by Selting [1]. We continue this line of research with a *quantitative* analysis of the prosody of incomplete holds in comparison to complete turn holds and turn changes in Austrian German.

Sacks [3] and Selting [1] argued that if speakers want to hold their turn at the end of a TCU, they have to employ specific turn holding strategies, such as rush-through or a specific prosody that is distinct from the prosody of a turn change. In incomplete holds, turn-holding is already signalled by syntactic projection. Therefore, we expect a prosodic realisation of incomplete holds that is distinct from both complete holds and turn changes. We draw the following hypotheses from the literature:

Final intonation. Ford & Thompson [4] reported a considerable number of syntactically complete turn holds that are also prosodically complete instead of projecting more talk by means of prosody. Incomplete holds, on the other hand, are expected to display prosody that does project more talk. Therefore, we expect more continuing intonation contours in incomplete holds, more terminal contours in change, and terminal as well as continuing contours in complete holds.

Final F0. Turn changes have been found to be associated with rising or falling pitch at the end [5, 6, 7] while turn holds displayed a flat pitch in the middle of a speaker's range [1, 6, 7, 8, 9]. A study of Swedish, on the other hand, found that while level pitch was a turn-holding cue, rising pitch was not clearly associated with either hold or change [10]. However, most of these studies did not distinguish syntactically complete from incomplete holds. Incomplete holds have been described to display final level or (slightly) rising pitch [1, 11]. Thus, we expect to find flat F0 in incomplete holds.

Intensity. We expect turn changes to have lower intensity than turn holds [9]. Within holds, we expect higher intensity in complete holds than in incomplete holds due to the need for specific prosodic turn-holding devices [1, 3], which might not be necessary when more talk is projected already by the incomplete syntax.

Articulation rate. Findings for lengthening are not consistent. Studies found more final lengthening in change than in hold [5], the reverse pattern [6, 7] or no consistent pattern at all [12]. For incomplete holds, Local & Kelly [11] found no noticeable slowing of tempo at the end. In the data analysed for this study, however, we had the auditory impression that speakers frequently do slow down in incomplete holds. Thus, we expect a lower articulation rate in incomplete hold, but we do not have clear expectations about the articulation rate in complete hold

*equal contribution

vs. change.

To address these hypotheses, we analyse the data two steps. First, we perform a Random Forest (RF) classification. The purpose of the RF is not classification per se, but to tell us *which* features are important for distinguishing the three turn-taking categories. Second, we perform regression analyses to find out *how* exactly the features are related to incomplete hold, complete hold and change.

2. MATERIALS

This study is based on 70 minutes of conversation (5 minutes in 14 conversations: 14f & 14m speakers) from the *Graz Corpus of Read and Spontaneous Speech* (GRASS) [13]. The corpus contains spontaneous face-to-face conversations between pairs of native Austrian German speakers. The data was annotated on three levels: turn-taking, prosodic phrases, and a phonetic segmentation via forced alignment, which was manually corrected [14].

For both prosodic phrasing and turn-taking, annotations were created in two stages: One annotator created a first version, which was then corrected by one of the other annotators. Prosodic phrases were labelled as *termination*, *continuation* or *high-rise* [15]. In Schuppler et al. [15], *continuation* was called *rise*, but other intonation contours indicating continuation were also annotated with this label.

Turn-taking was labelled in Inter-Pausal-Units (IPUs) separated by pauses longer than 150 ms [16]. IPUs were labelled as `change` before a turn change (questions are not included); as `complete hold` for a syntactically complete turn hold; and as `incomplete hold` if the same speaker continued talking after producing a pause at a point of *maximum grammatical control* [2]. The labelling process was based on sequential criteria in the tradition of Conversation Analysis (i.e., on interlocutors' behaviours and not on a speaker's intentions).

From a set of $N = 1011$ tokens, we excluded tokens containing laughter, uncertain labels and turn-changes with overlapping speech at the end of the IPU, resulting in $N = 837$ tokens (`in-hold`: 222, `com-hold`: 368, `change`: 247) for this study.

3. METHODS

3.1. Features

For each target IPU, we extracted 10 acoustic features (5 F0, 4 intensity, 1 duration) and the categorical feature *phrase-final intonation*. Since we are interested in turn-holding and turn-yielding practices, acoustic features were calculated for a window at the end of each IPU. When comparing the classification performance (cf., sec. 3.2) of three different window lengths (0.6, 0.8 and 1.0 seconds), a window of 0.6 s yielded the best performance (mean F_1-scores over all turn-taking categories: $F_{1,0.6\,\mathrm{s}} = 0.61$, $F_{1,0.8\,\mathrm{s}} = 0.60$, $F_{1,1.0\,\mathrm{s}} = 0.60$). If an IPU was shorter than 0.6 s, features were calculated for this shorter window.

F0 was tracked with the Python [17] package Parselmouth [18], and corrected for octave jumps with [19]. F0 was speaker-normalised by converting Hertz to semitones based on each speaker's overall median F0; then we calculated the maximum, minimum ($\mathrm{F0_{max}}$, $\mathrm{F0_{min}}$), median and range ($\mathrm{F0_{med}}$, $\mathrm{F0_{range}}$) and the $\mathrm{F0_{slope}} = (\Delta \mathrm{F0})/(\Delta t_{(\mathrm{F0_{min}},\,\mathrm{F0_{max}})})$.

We calculated the articulation rate (*ArtR*), and four intensity features ($\mathrm{I_{max}}$, $\mathrm{I_{med}}$, $\mathrm{I_{std}}$ and the position in time of $\mathrm{I_{max}}$ ($t(\mathrm{I_{max}})$)) as z-score speaker-normalised root mean square (RMS) values.

finIntonation indicates the annotated (functional) phrase-final intonation contour (levels: terminal, continuing, high-rising contour). It is important to note that these annotations are not the same as the final F0 slope; for instance, continuing intonation subsumes various final F0 movements (cf., comma intonation in [20]).

3.2. Random forest classifier

We used a random forest (RF) classifier from *scikit-learn* [21] (with 100 estimators, the square root as maximum number of features and Gini impurity) that we trained on the 11 features described in section 3.1. We trained the classifier for three different comparisons: `in-hold` vs `com-hold`, `in-hold` vs. `change` and `com-hold` vs. `change`. For each comparison, we split the data into 80 % training and 20 % test data and cross-validated with ten randomly chosen splits. Cross-validation is crucial since our data shows much variation both between and within speakers. With these pairwise comparisons, we obtained a better classification and thus a more reliable feature ranking than with a three-way classification.

3.3. Regression models

We built linear mixed effects regression models with R's [22] `lme4` package [23], where the dependent variable was one of the 10 continuous acoustic features. Models included the independent variables *Category* (`in-hold`, `com-hold`, `change`), and *Speaker* ($N = 18$) as a random variable. We then performed pairwise comparisons of the three values of the variable *Category* with `emmeans` [24]. For the categorical variable *finIntonation*, we built a logistic

in-hold vs. com-hold	in-hold vs. change	com-hold vs. change
$F_{1,\texttt{in-hold}}$=.60	$F_{1,\texttt{in-hold}}$=.78	$F_{1,\texttt{com-hold}}$=.68
$F_{1,\texttt{com-hold}}$=.79	$F_{1,\texttt{change}}$=.81	$F_{1,\texttt{change}}$=.40

Table 1: F_1 scores of the three RF classifications.

regression model with `multinom` from `nnet` [25], with *finIntonation* as dependent variable and *Category* as independent variable. We rotated the reference levels of *finIntonation* and *Category* to examine all comparisons. Here, we present significant predictors only. Features and model outputs can be found at[1].

4. RANDOM FOREST RESULTS

Table 1 presents F1 scores of the pairwise classifications. Two of the three pairwise comparisons yielded good results in the RF classification: `in-hold` vs. `com-hold` and `in-hold` vs. `change`. The classification of `com-hold` vs. `change` showed a bad performance. `in-holds` could be distinguished correctly from `changes` in 75.45% of cases (true positives), but from `com-holds` only in 56.36% of cases. `com-holds` were correctly distinguished from `in-holds` in 82.03% of cases, and from `changes` in 74.73% of cases. `changes` were classified correctly in 83.60% of the cases when compared to `in-holds` but only in 34.49% when compared to `com-holds`.

Table 2 shows the four highest ranked features for the pairwise comparisons with their corresponding importance values. We do not present further ranks because of their low importance values. *finIntonation* was the most important feature in distinguishing `in-hold` from the other two categories. Due to the low importance values of the other features and the overall poor classification performance of `com-hold` vs. `change`, the ranking of these features should be taken with caution.

5. REGRESSION RESULTS & DISCUSSION

5.1. Final intonation

Figure 1 presents the distribution of the annotated phrase-final intonation contours over the three turn-taking categories. The logistic regression showed that in `change`, terminal intonation was significantly more likely than continuing ($p < .0001$) and high-rising intonation ($p < .05$), compared to `com-hold`. In `change`, terminal intonation was significantly more likely than continuing ($p < .0001$) and high-rising intonation ($p < .01$), compared to `in-hold`. In `in-hold`, continuing intonation was more likely

in-hold vs. com-hold		in-hold vs. change		com-hold vs. change	
finInt	.123	finInt	.222	I_{med}	.021
ArtR	.012	*ArtR*	.024	*ArtR*	.015
I_{med}	.009	I_{med}	.008	I_{max}	.014
$t(I_{max})$	.007	I_{std}	.008	$F0_{slope}$	.014

Table 2: Highest ranked features with respective importance values of the three RF classifications – a higher value indicates a higher importance.

than high-rise ($p < .05$), compared to `change`. In `in-hold`, continuing intonation was more likely than high-rising ($p < .001$) and terminal intonation ($p < .0001$), compared to `com-hold`. Though the majority of `com-holds` had a terminal contour, the models do not indicate that this relationship is significant. The relatively high percentage of terminal intonation in `com-holds` (49% of all terminal contours were produced in `com-holds`) fits with the finding that only about half of all transition relevance places actually involve a turn-shift [4].

As expected, `in-holds` are clearly associated with continuing intonation and `changes` with terminal intonation, and *finIntonation* was the most important feature in the RF feature ranking (except for `com-hold` vs. `change`). High-rises occurred most frequently in `com-holds`, even though they were generally rare, but `com-holds` were not consistently associated with either terminal or continuing intonation, which also matches our expectations.

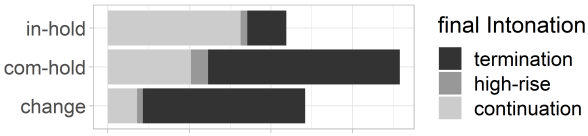

Figure 1: Distribution of phrase-final intonation labels over the three turn-taking categories.

5.2. Final F0

$F0_{max}$ was higher for `com-hold` than for `change` ($p < .05$) as well as `in-hold` ($p < .05$), but `in-hold` vs. `change` was not significant. $F0_{range}$ was higher for `com-hold` than `change` ($p < .05$) as well as `in-hold` ($p < .01$), but `in-hold` vs. `change` was not significant. $F0_{med}$ was higher for `com-hold` than for `change` ($p < .05$). For $F0_{min}$ and $F0_{slope}$, none of the comparisons showed a significant difference between the turn-taking categories. Figure 2 shows the distributions of $F0_{slope}$ values over the three categories. The two modes in the bimodal distribution indicate falling and rising contours. Even though distributions overlap considerably, some tendencies are visible: In `in-hold`, the peak at negative val-

ues is considerably higher than the one at positive values, indicating more falls than rises. The peak at negative values is close to zero, and the density around zero is clearly higher than in the other categories. This indicates that falls are not as steep and there are more flat F0 curves in `in-holds`. In `change`, the peak at negative values is much higher than the one at positive values, and the area under the curve at negative values extends more towards lower values, indicating that `changes` are more often falling than rising and that they display steeper falls than the other categories. In `com-hold`, the peak at positive values is only slightly higher than the one at negative values, but the area under the curve extends more towards higher values, indicating that, even though there are also falls in `com-holds`, there are overall more rises, and the rises are steeper than in other categories. Thus, we found more flat slopes as well as narrower F0 ranges in `in-holds`, which meets our expectations, more and steeper falls in `changes`, and more and steeper rises as well as higher F0 and a larger F0 range in `com-holds`. The higher F0 values in `com-holds` could be related to the higher percentage of high-rises in this category. However, even though some of the F0 features were significant, they were generally of low importance in the RF, except for the F0 slope in `change` vs. `com-hold`.

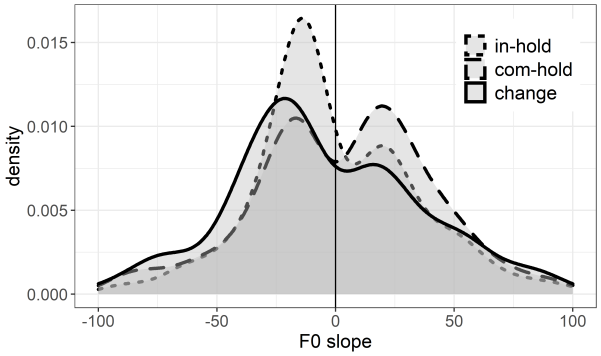

Figure 2: Density plot of $F0_{slope}$ for `in-hold` (dotted), `com-hold` (dashed) and `change` (solid). Positive values indicate a rise, negative values a fall, and values around zero indicate a flat F0.

5.3. Intensity

I_{med} was higher for `in-hold` than for `change` ($p < .001$) and `com-hold` ($p < .01$). I_{max} was marginally significantly higher for `in-hold` than for `change` ($p = .092$), but the other comparisons were not significant. I_{std} and the position of I_{max} were not significantly affected by the turn-taking categories. Thus, we did not find the expected higher intensity in `com-hold`. Instead, we found higher intensity in `in-hold`. This higher intensity at the end of `in-holds` might be due to higher sub-glottal pressure in the middle than at the end of a TCU, or it might be a strategy to contrast them to trail-offs [26].

5.4. Articulation rate

ArtR was highest for `change`, intermediary for `com-hold` and lowest for `in-hold` (`com-hold` vs. `change`: $p < .01$; `in-hold` vs. `change`: $p < .0001$; `com-hold` vs. `in-hold`: $p < .0001$). Thus, we found the fastest articulation rate in `change`, and, as expected, the slowest articulation rate in `in-holds`. This result does not confirm Local & Kelly's [11] observation of no noticeable slowing, but could be interpreted in the light of hesitations or disfluency. While turn-taking was not explicitly investigated by Betz et al. [27], they found disfluent lengthening most often in cases which could be characterised as points of *maximum grammatical control* (e.g., in articles, conjunctions and prepositions).

6. CONCLUSION

In this study, we analysed the prosody of syntactically incomplete turn-holds, that is, when a pause is produced at a point of *maximum grammatical control*, and compared them to syntactically complete turn-holds as well as to turn-changes. Our analyses showed that syntactically complete holds stand out through a higher maximum and median F0, which is probably due to a higher percentage of high-rises in this category. However, contrary to the expectation of more prosodic effort to distinguish syntactically complete holds from syntactically complete changes [1, 3], we did not find a higher intensity in complete holds, a specific articulation rate or a clear association with continuing intonation. Turn-changes, on the other hand, are characterised by terminal intonation and a higher articulation rate. The poor RF classification of syntactically complete turn-holds vs changes, suggests an insufficient prosodic distinction between these two categories. Overall, we found that the prosody at points of *maximum grammatical control*, that is, at the end of syntactically incomplete holds, is characterised by continuing intonation, high intensity, low articulation rate, and a flatter F0 than grammatically complete holds.

7. ACKNOWLEDGEMENTS

The work by A. Kelterer and S. Wepner was funded by grant P-32700-NB from FWF (Austrian Science Fund).

8. REFERENCES

[1] Selting, M. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics* 6(3), 371–388.

[2] Schegloff, E. A. 1996. Turn organization: One intersection of grammar and interaction. In: Ochs, E., Schegloff, E. A., Thompson, S. (eds), *Interaction and grammar*. Cambridge: Cambridge University Press, 52–133.

[3] Sacks, H., Schegloff, E., Jefferson, G. 1974. A simplest systematics for the organisation of turn-taking for conversation. *Language* 50, 696–735.

[4] Ford, C., Thompson, S. 1996. *Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns*, 134–184. Cambridge University Press.

[5] Local, J., Kelly, J., Wells, W. 1986. Towards a phonology of conversation: turn-taking in Tyneside English. *J. Linguist.* 22(2), 411–437.

[6] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech* 41, 295–321.

[7] Gravano, A., Hirschberg, J. 2011. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang* 25(3), 601–634.

[8] Duncan, S. 1972. Some signals and rules for taking speaking turns in conversations. *J.Pers.Soc.Psychol.* 23(2), 283–292.

[9] Skantze, G. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67, 1–26.

[10] Edlund, J., Heldner, M. 2005. Exploring prosody in interaction control. *Phonetica* 62(2-4), 215–226.

[11] Local, J., Kelly, J. 1986. Projection and 'Silences': Notes on Phonetic and Conversational Structure. *Human Studies* 9(2/3), 185–204.

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

[13] Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., Pessentheiner, H. 2014. GRASS: The Graz corpus of Read And Spontaneous Speech. *Proceedings of LREC*, 1465–1470.

[14] Ludusan, B., Schuppler, B. 2022. An analysis of prosodic boundaries across speaking styles in two varieties of german. *Speech Communication*.

[15] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94, 62–74.

[16] Schuppler, B., Kelterer, A. 2021. Developing an Annotation System for Communicative Functions for a Cross-Layer ASR System. *Proceedings of Workshop "Integrating Perspectives on Discourse Annotation"*.

[17] Python Software Foundation, 2009. Python Language Reference. Version 3.9.

[18] Jadoul, Y., Thompson, B., de Boer, B. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71, 1–15.

[19] Paierl, M. 2021. Detektion und Visualisierung von F0-Berechnungsfehlern.

[20] Chafe, W. 1988. Linking intonation units in spoken english. In: Haiman, J., Thompson, S. A. (eds), *Clause combining in grammar and discourse*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 1–27.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

[22] R Core Team, 2020. R: A Language and Environment for Statistical Computing. Version 4.2.1.

[23] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[24] Lenth, R. 2020. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1.

[25] Venables, W. N., Ripley, B. D. 2002. *Modern Applied Statistics with S*. New York: Springer fourth edition. ISBN 0-387-95457-0.

[26] Walker, G. 2012. Coordination and interpretation of vocal and visible resources: 'Trail-off' conjunctions. *Language and Speech* 55(1), 141–163.

[27] Betz, S., Wagner, P., Voße, J. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. *Tagungsband der 12. Tagung Phonetik und Phonologie im Deutschsprachigen Raum*.

[1] Feature table and model outputs can be found at: https://cloud.tugraz.at/index.php/s/mAH4LC2NadQcXSi