PITCH SLOPE AND END POINT AS TURN-TAKING CUES IN SWEDISH

Mattias Heldner & Marcin Włodarczak

Department of Linguistics, Stockholm University, Sweden
heldner@ling.su.se, wlodarczak@ling.su.se

ABSTRACT

This paper examines the relevance of parameters related to slope and end-point of pitch segments for indicating turn-taking intentions in Swedish. Perceptually motivated stylization in Prosogram was used to characterize the last pitch segment in talkspurts involved in floor-keeping and turn-yielding events. The results suggest a limited contribution of pitch pattern direction and position of its endpoint in the speaker's pitch range to signaling turn-taking intentions in Swedish.

Keywords: turn-taking cues, spontaneous dialogue, pitch stylization, Prosogram.

1. INTRODUCTION

This paper revisits the topic of pitch patterns as potential turn-taking cues in Swedish. Previous work for Swedish, as well as for several other languages, generally associates static (i.e. flat) pitch segments with floor-keeping and dynamic (i.e. rising or falling) pitch segments with turn-yielding, see for example [3, 5, 7, 11, 14, 16, 19, 26, 28, 29]. Studies using more fine-grained models of intonation than just rising, falling and flat intonation typically also include slowly rising and slowly falling segments among the floor-keeping cues [12, 27]. In addition, they indicate that the endpoint of the pitch segment is relevant and associate pitch segments reaching the top or bottom of the speaker's pitch range with turn-yielding and those ending somewhere in the middle with floor-keeping [12, 27]. However, prior research on pitch patterns as turn-taking cues has predominantly consisted in either qualitative descriptions of intonation categories on relatively limited materials, or quantitative studies using fairly ad-hoc criteria for the static versus dynamic pitch distinction. In this study, we aim at combining perceptually motivated analytic categories with large-scale corpus work by using Prosogram [4, 20].

Prosogram attempts to provide a representation of intonation in speech as it is *perceived* by human listeners, using a model of tonal perception. Briefly, Prosogram characterizes pitch patterns in syllable nuclei as either static (i.e. flat) or dynamic (i.e. rising or falling) linear pitch segments using an auditory threshold for detection of pitch movement. This 'glissando threshold' is dependent on the square of the duration of the stylized pitch segment. Furthermore, a 'differential glissando threshold' is used to detect changes in slope within the syllable nuclei, allowing for rising-falling or falling-rising patterns, for example. Prosogram includes automatic segmentation into units approximating syllabic nuclei and does not require any manual labeling. Additionally, Prosogram estimates a number of more global prosodic parameters including speaker pitch range. Prosogram is implemented as a Praat script and is available under a Creative Commons attribution and a non-commercial use license [23].

For our purposes, Prosogram has several desirable properties. First of all, it transforms acoustic fundamental frequency curves into an approximation of perceived pitch patterns in an objective and repeatable way. Second, it includes a straightforward distinction between static and dynamic segments, and characterizes all segments such that gradual slopes as well as start and end points can be extracted. Third, the stylization into linear pitch segments eliminates the need for other techniques for data reduction (e.g. fitting regression lines or splines) that would otherwise have been necessary for quantitative treatment of the data. Fourth, the speaker pitch range estimations allow a sensible speaker normalization of extracted data points. These advantages have been acknowledged in a number of previous studies [17, 18, 25].

In this work, we apply Prosogram to a relatively large Swedish speech material, a subset of the Spontal corpus [6] for which events related to floor-keeping and turn-yielding have been automatically annotated [13]. We extract parameters characterizing the last voiced segment in each talkspurt from the Prosogram-stylized pitch patterns and examine the distributions of these parameters across floor-keeping and turn-yielding events. Finally, we use logistic regression to examine the relevance of parameters related to slope and end-point of the pitch segments for indicating turn-taking intentions. In particular, we test the hypotheses 1) that floor-keeping is associated with static pitch segments and turn-yielding with dynamic ones, and 2) that floor-keeping is associated with pitch segments ending in the middle of the speaker's pitch range and turn-yielding with segments reaching the top or bottom of the speaker's pitch range.

2. METHOD AND MATERIAL

2.1. The Spontal Corpus

The speech material used in this work was drawn from the Spontal corpus [6]. Spontal consists of recordings of audio, video, and three-dimensional motion capture from around 120 half-hour sessions of spontaneous two-party face-to-face conversations in Swedish. Here, we used the close-talk microphone recordings from a subset of this corpus (24 dialogues), for a total of 12 hours and 58 minutes of recordings. This subset includes data from 19 female and 29 male speakers, and contains 4 female-female dyads, 11 female-male dyads and 9 male-male dyads.

2.2. Annotation of floor-keeping and turn-yielding

The speech material had previously been automatically annotated for events related to floor-keeping and turn-yielding [13] using a computational model of interaction closely resembling that in [15]. Briefly, this method used speech activity detection from each dyad to identify a number of dialogue states: intervals of single-speaker speech for each speaker; intervals of joint silence; and intervals of joint speech. The sequence of such dialogue states was subsequently used to identify floor-keeping events (defined as joint silences preceded and followed by single-speaker speech by the same speaker) and turn-yielding events (defined as joint silences or joint speech preceded and followed by single-speaker speech by different speakers). Sequences of joint speech preceded and followed by single-speaker speech by the same speaker were also identified so that they could be excluded from further analyses. This automatic annotation has recently been incorporated in TextGridTools [2]. In addition, all talkspurts were subdivided into very short utterances (VSUs) and their complement (NonVSUs) based on their duration. Talkspurts between 200 ms to 1000 ms were labeled VSUs and Talkspurts longer than 1000 ms were labeled NonVSUs [8, 9].

2.3. Prosogram

We used a slightly modified version of Piet Mertens Prosogram (version 2.9f) to obtain stylized pitch patterns in the entire speech material [23]. Instead of using one of the provided segmentation methods, we modified the script to stylize entire voiced intervals (as determined by the voiced/unvoiced distinction) rather than syllable nuclei within voiced intervals. This modification was relevant here, as we wanted to capture talkspurt-final pitch movements irrespective of whether they occurred in vowels or in syllable rhymes. We used different pitch detection ranges for males (60 to 300 Hz) and females (100 to 450 Hz). The following Prosogram parameter settings were used: Frame period 0.005 s; Glissando threshold $G = 0.16/T^2$ semitones per second (ST/s) where T is the duration of the tonal segment; Differential glissando threshold DG = 20 ST/s, minimum duration of tonal segments dmin = 0.035 s. The modification and choice of parameter settings both contributed to making stylization more sensitive to dynamic segments than the default settings of Prosogram.

2.4. Local and global pitch features

A Java program was used to extract an initial set of local pitch features from the stylized tonal segments produced by Prosogram at locations determined by the annotations of turn-taking events. These local pitch features included, among other things, the number of sub-segments in the last tonal segment in the talkspurt (in case the differential glissando threshold was exceeded); the start and end times for all sub-segments of the last tonal segment in the talkspurt; and stylized pitch values (in Hz) at the start and end times of all tonal sub-segments in the talkspurt.

In addition, a set of more global pitch features related to speaker pitch range were extracted from the Prosogram output. These features included bottom, mean, median, and top (in ST relative to 1 Hz), and pitch range (in ST). The pitch range was defined as the distance between bottom and top lines, which in turn were defined as the 2nd and 98th percentiles of all stylized tonal segments from each speaker.

All local and global pitch features were gathered in a “big table” and a set of additional features were calculated. All pitch values expressed in Hz were transformed into ST relative to 1 Hz. Duration, pitch difference and pitch slope (of all tonal sub-segments in the last tonal segment in the talkspurt) were calculated. Pitch slope was expressed in ST per decisecond (1 ds = 0.1 s) in order better to reflect the degree of pitch change occurring within a unit approximating the duration of a vowel. The pitch start and end points were normalized to the pitch range of the speaker by subtracting the value of the median line (i.e. they were expressed as distance in ST relative to the speaker’s median pitch).

2.5. Preprocessing

The extraction resulted in pitch features from 12839 talkspurts. However, more than half of these were excluded from further analyses for a couple of reasons. First, 4688 VSUs were excluded, as many VSUs are backchannel-like utterances that are linked to qualitatively different floor-negotiation strategies [8]. Next, 1375 'outlier' talkspurts were excluded from further analysis, either because they contained tonal segments outside the speaker's pitch range, or tonal segments ending earlier than 500 ms before the talkspurt offset. Finally, the last voiced segment of these talkspurts contained up to 13 tonal sub-segments. Due to space limitations, we excluded 833 talkspurts, and will only present results from talkspurts with one or two sub-segments here.

3. RESULTS

5943 talkspurts remained after the preprocessing. These included 2571 floor-keeping talkspurts (43.3%), and 3372 turn-yielding talkspurts (56.7%). 2024 instances of the turn-yielding talkspurts were followed by joint silence, while the speaker switch occurred in overlap in the remaining 1348.

3.1. Distribution of pitch contours

Table 1 shows the distribution of pitch contours in the last tonal segment of talkspurts (with one or two sub-segments) across floor-keeping and turn-yielding events. Anything that deviates from static pitch in this table is labeled either fall or rise, irrespective of the angle of inclination. This analysis revealed, unsurprisingly, that a large proportion of all pitch contours (45.8%) had static, or in other words perceptually flat pitch according to the Prosogram stylization. Among the segments classified as dynamic, that is, as perceptually rising or falling according to Prosogram, falling pitch (here defined as 1 or 2 sub-segment stylizations ending in falling pitch) was considerably more frequent (40.2%) than rising pitch (13.9%).

Table 1: Distribution of pitch contours in the last tonal segment of a talkspurt (with one or two sub-segments) across turn-taking events.

Pitch contour	Floor-keeping	Turn-yielding	Total
flat	1075	1649	2724
fall	801	843	1644
fall-fall	126	109	235
flat-fall	135	145	280
rise-fall	95	138	233
rise	146	204	350
rise-rise	7	16	23
flat-rise	40	82	122
fall-rise	146	186	332
Total	2571	3372	5943

Furthermore, Table 1 clearly demonstrates that the direction of the last pitch segment is not sufficient to determine whether the talkspurt will be followed by more speech by the same speaker, or by a speaker switch. Frequencies of specific pitch contour types are comparable within the two turn-taking events. Flat pitch is frequent in both types of turn-taking events, but contrary to the expectations relatively more frequent among the turn-yielding cases (48.9%) than the floor-keeping ones (41.8%). Furthermore, falling pitch is, surprisingly enough, relatively more frequent in floor-keeping (45.0%) than in turn-yielding (36.6%). Rising pitch, finally, is nearly equally frequent in turn-yielding (14.5%) and floor-keeping (13.2%) talkspurts.

3.2. Logistic regression

To examine the importance of pitch features for indicating turn-taking intentions, we used logistic regression. Specifically, to test the hypotheses that distance from mid-pitch range and divergence from flat pitch predict speaker change, we used the absolute pitch slope (i.e. no distinction between falls and rises); the absolute distance between the end-point of the last pitch segment and the median of the speaker's pitch range; and the interaction between these as predictors of the binary outcome variable turn-yielding vs. floor-keeping.

We followed the process of fitting a logistic regression model outlined in [10]. First we ran an initial hierarchical analysis where the logistic regression model was built up one predictor at a time to ascertain whether the 'new' model improved the fit compared to the previous model. This analysis revealed that both main effects made significant contributions to the model, both according to changes in likelihood ratio statistics and according to the Wald statistic, whereas the two-way interaction was not significant. Subsequently, we specified a final model including only the significant predictors and re-ran the analysis to save diagnostic statistics and check the residuals. We also ran a bootstrapped model to obtain 95% bootstrap confidence intervals. The results of these analyses are shown in Table 2.

The 95% confidence intervals for the regression coefficient B showed that there is indeed a genuine and positive relationship between the distance from mid-pitch range and the probability of a speaker change. The same holds for the divergence from flat pitch. Thus, if the distance between endpoint and median increases, or if the absolute slope increases, so will the probability of a speaker change.

Table 2. Coefficients of the model predicting whether a talkspurt was followed by a speaker change. 95% bootstrap confidence intervals for B based on 1000 samples. Note. $R^2 = 0.004$ (Hosmer & Lemeshow); 0.006 (Cox & Snell); 0.007 (Nagelkerke). Model $\chi^2(2) = 32.831, p < .01.$

	95% CI for B						95% CI for Odds Ratio		
	Lower	B	Upper	S.E.	Wald	Sig.	Lower	Odds	Upper
Distance from mid-pitch	0.023	0.044	0.067	0.011	14.077	0.000	1.021	1.045	1.069
Divergence from flat pitch	0.019	0.039	0.061	0.011	14.195	0.000	1.019	1.040	1.061
Constant	0.033	0.110	0.187	0.039	7.814	0.005		1.117	

However, an inspection of the odds ratios in Table 2 will show that the effects are marginal. For example, if the distance from mid-pitch increases by 1 ST, the probability of a speaker change will increase by 4.5%, and if the slope increases by 1 ST/ds, the probability of a speaker change will increase by 4%. Similarly, the different R^2 measures (indicating how well the model fits the data) indicate that although the predictors were significant, the model explained only a negligible fraction of the variance.

4. DISCUSSION AND CONCLUSIONS

To summarize, with features extracted from the last voiced segment of the talkspurt, neither categorical pitch patterns related to direction of pitch segments, nor gradual pitch features related to divergence from flat pitch and distance from the middle of the speaker's pitch range appear to be particularly strong indicators of turn-taking intentions in Swedish. It does not seem to be the case that flat pitch is clearly associated with floor-keeping and that rising or falling pitch is associated with turn-yielding. Neither are pitch segments ending somewhere in the middle of the speaker's pitch range clearly associated with floor-keeping and pitch segments ending in the periphery of the pitch range with turn-yielding. There were significant positive relationships, but the effect sizes were on the whole tiny.

These results could of course be related to the methods and speech materials used here. For example, the large proportion of flat pitch segments was to some extent expected given the categorical static vs. gradual dynamic description of pitch direction in Prosogram. However, the different proportions of rises and falls cannot be attributed to Prosogram—rather they must be related to the spontaneous nature of the dialogues and/or the tonal characteristics of Swedish. Furthermore, several authors have pointed out the optionality of all turn-taking decisions, which presents problems for automatic annotation of turn-taking events [12, 27]. Thus, the large proportion of flat pitch segments in turn-yielding events could in principle be due to situations where the current speaker intended to continue, but where the next speaker grabbed the floor during the pause. This is indeed a common interruption strategy in spontaneous interaction [24]. It is worth noting that such situations would be difficult to handle in any type of annotation.

With respect to the inferential statistics, previous studies reporting positive relationships between intonation categories and turn-taking intentions [e.g. 27] used discrete intonation categories related to slope and end point in range, whereas we used gradual pitch features. There is the possibility that we would have obtained larger effect sizes with categorical units related to slope and end point in range. But importantly, there is also the possibility that *the last voiced segment* in the talkspurt provide little or no information relevant for indicating turn-taking intentions. We are not yet ready to abandon the idea that intonation provides information that is relevant for indicating turn-taking intentions. In future work, we would like to explore relations between the last and the last but one of the voiced segments. That is, pitch relations between consecutive syllables rather than just the shape of the final syllable.

On a different note, leaning on the experiences gained in this study, we argue that Prosogram is a considerably better basis for various symbolic or phonological representations of intonation than the raw fundamental frequency contour. In future work, we will continue exploring Prosogram for identifying candidate pitch contours underlying symbolic tonal inventories of Swedish. This is a particularly promising line of extending its limited set of boundary tones [1]. Indeed, Prosogram has recently been used as a basis for automatic symbolic transcription of prosody [21, 22]. We hope to try out this symbolic transcription of prosody on our Swedish material.

7. ACKNOWLEDGMENTS

This work was partly funded by the Swedish Research Council grant 2014-1072 *Breathing in conversation*. We thank Anna Hjalmarsson, Sofia Strömbergsson and Jens Edlund for many useful discussions.

8. REFERENCES

[1] Bruce, G. 2005. Intonational prominence in varieties of Swedish revisited. In: Jun, S.-A. (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 410-429.

[2] Buschmeier, H., Włodarczak, M. 2013. TextGridTools: A TextGrid processing and analysis toolkit for Python. In: *Proc. 24 Konferenz zur Elektronischen Sprachsignalverarbeitung*. Bielefeld, 152–157.

[3] Caspers, J. 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics* 31, 251–276.

[4] d' Alessandro, C., Mertens, P. 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9, 257-288.

[5] Duncan, S., Jr. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23, 283-292.

[6] 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: *Proc. LREC 2010*. Valetta, 2992-2995.

[7] Edlund, J., Heldner, M. 2005. Exploring prosody in interaction control. *Phonetica* 62, 215-226.

[8] Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., Hirschberg, J. 2010. Very short utterances in conversation. In: *Proc. Fonetik 2010*. Lund, 11-16.

[9] Edlund, J., Heldner, M., Pelcé, A. 2009. Prosodic features of very short utterances in dialogue. In: *Nordic Prosody: Proc. of the Xth Conference, Helsinki 2008*. Frankfurt am Main: Peter Lang, 57-68.

[10] Field, A. 2013. *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll.* Los Angeles: Sage.

[11] Ford, C., Thompson, S. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In: Ochs, E., Schegloff, E., Thomson, S. (eds.), *Interaction and grammar*. Cambridge: Cambridge University Press, 134-184.

[12] Gravano, A., Hirschberg, J. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language* 25, 601-634.

[13] Heldner, M., Edlund, J., Hjalmarsson, A., Laskowski, K. 2011. Very short utterances and timing in turn-taking. In: *Proc. Interspeech 2011*. Florence, 2837-2840.

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

[15] Jaffe, J., Feldstein, S. 1970. *Rhythms of dialogue.* New York: Academic Press.

[16] Jefferson, G. 1984. Transcript notation. In: Atkinson, M., Heritage, J. (eds.), *Structure of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, ix-xvi.

[17] Karpiński, M. 2011/12. Acoustic properties and functions of phrase-final rises in Polish task-oriented dialogues. In: *Speech and Language Technology*. Poznan: Polish Phonetic Association, 147-156.

[18] Karpiński, M., Szalkowska-Kim, E. 2011/12. On intonation of questions in Korean and Polish task-oriented dialogues. In: *Speech and Language Technology*. Poznan: Polish Phonetic Association, 137-147.

[19] 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.

[20] Mertens, P. 2004. The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In: *Proc. Speech Prosody 2004*. Nara, 549-552.

[21] Mertens, P. 2013. Automatic labelling of pitch levels and pitch movements in speech corpora. In: *Proc. TRASP 2013*. Aix-en-Provence, 42-46.

[22] Mertens, P. 2013. From pitch stylization to automatic tonal annotation of speech corpora. In: *Rhapsodie: A prosodic and syntactic treebank for spoken French*. Amsterdam: Benjamins.

[23] Mertens, P. 2015. Prosogram (v2.10). Retrieved from http://bach.arts.kuleuven.be/pmertens/prosogram/

[24] Özgür, Ç., Shriberg, E. 2006. Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site. In: *Machine Learning for Multimodal Interaction*. Berlin: Springer, 212-224.

[25] Patel, A. D., Iversen, J. R., Rosenberg, J. C. 2006. Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America* 119, 3034-3047.

[26] 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.

[27] Wennerstrom, A., Siegel, A. F. 2003. Keeping the floor in multiparty conversations: Intonation, syntax, and pause *Discourse Processes* 36, 77-107.

[28] Yanushevskaya, I., Kane, J., De Looze, C. l., Ní Chasaide, A. 2014. The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps. In: *Proc. Speech Prosody 7*. Dublin, 959-963.

[29] Zellers, M. 2013. Pitch and lengthening as cues to turn transition in Swedish. In: *Proc. Interspeech 2013*. Lyon, 248-252.