VOICE QUALITY DYNAMICS OF TURN-TAKING EVENTS IN SWEDISH AND GERMAN

Marcin Włodarczak*, Mattias Heldner*, Anna Bruggeman†, Petra Wagner†

*Stockholm University, †Bielefeld Univeristy
{wlodarczak,heldner}@ling.su.se, {anna.bruggeman,petra.wagner}@uni-bielefeld.de

ABSTRACT

The study investigates variation of two voice quality features (smoothed cepstral peak prominence and the relative strength of the fundamental) preceding speaker changes and turn holds in spontaneous conversations in Swedish and German. We demonstrate that the overall pattern is to a large extent linked to the syllabic structure in speech. While speech before turn holds is characterised by more periodic phonation than before speaker changes, the effect sizes are very small and are unlikely to be of practical importance for speech communication. We observe no evidence in favour of language specificity of the reported contrasts.

Keywords: phonation type, turn-taking, neck-surface acceleration, Generalized Additive Mixed Models

1. INTRODUCTION

The speech signal is rich in linguistic and paralinguistic information [1], including prosodic patterns related to coordination of turn-taking in conversation [2, 3]. However, studies of prosodic turn-taking cues rarely include properties related to voice quality (i.e. phonation type), in spite of a growing body of evidence for the importance of voice control to prosodic expression [4, 5, 6]. This state of affairs can to a large extent be explained by the difficulty of studying the voice source in spontaneous speech, where inverse filtering of supraglottal resonances is either prohibitively time consuming or error-prone [7].

Existing work has thus relied primarily on manual annotations of specific voice quality types (e.g. creak [8]) or acoustic analysis of the raw audio signal [9], often using features which are difficult to interpret when calculated from connected speech, such as jitter or shimmer [3, 10]. The methodological differences notwithstanding, this line of research has found that speech preceding speaker changes is less periodic and more creaky than in turn-holds.

More recently, [11] revisited the question of phonatory turn-taking cues using a signal collected with miniature accelerometers attached to speakers' tracheal wall below the larynx. The method is simple to use, noninvasive and can be used to estimate voice source parameters [12, 13, 14]. It has also showed promise for classification of voice quality [15], although it might offer little advantage over the microphone signal in highly controlled speech [16].

[11] used the method to analyze several voice quality parameters (harmonicity, strength of the fundamental and spectral balance) extracted from the last voiced interval preceding turn-taking events. They found that, in line with previous research, speaker changes accompanied by silence were characterised by less modal phonation than turn holds. In addition, they demonstrated that while some voice quality features did help to distinguish between the categories, their contributions were limited when compared to that of intensity level and, to a lesser extent, fundamental frequency (f_0).

An obvious limitation of the analysis in [11] is that it was based on averages calculated over the last voiced interval in a talkspurt. Thus, it failed to account for variability of the parameters in question over time. In this paper, we remedy this shortcoming by analysing the dynamical properties of voice quality using Generalized Additive Mixed Models (GAMM). By including two languages (German and Swedish), we address the question of language-specificity of voice quality variation.

2. METHOD

2.1. Material

The material consisted of seven dyadic conversations in Swedish and four in German. The participants (8 males and 14 females) were instructed to talk with each other on a topic of their choice for about 30 minutes (mean durations: 28 and 30 minutes for the German and Swedish conversations, respectively). All participants knew

each other before the recording. The recordings were made in sound treated rooms at Stockholm University and Bielefeld University. In addition to audio recordings using regular microphones, participants' vocal activity was also recorded with miniature accelerometers (Knowles BU-27135) attached to the neck below the level of the glottis.

2.2. Feature extraction

For Swedish, talkspurts were identified from the audio recordings using the voice activity segmentation method proposed in [17]. For German, due to strong crosstalk between the channels, talkspurts were labelled manually by a student assistant. Each silence was then automatically classified as either within- or between-speaker (henceforth, WSS and BSS), depending on whether it was associated with a speaker change (BSS) or a turn-hold (WSS). Given that very short utterances correspond largely to backchannels [18], WSS and BSS intervals in which the talkspurt preceding the silence was shorter than one second were excluded from the analysis.

The following voice quality features were extracted from voiced frames (50 ms in length with 2 ms step, Gaussian window) in the accelerometer signal:

Smoothed cepstral peak prominence (CPPS): amplitude of the first rahmonic relative to the regression line over the power cepstrum, in dB [19] characterising the periodicity of the signal.

***L_1L_2*:** characterising the strength of the fundamental relative to the second harmonic and/or spectral slope in the low frequency part of the spectrum: $L_1 - L_2$, in dB.

Given that voice source parameters vary significantly as a function of subglottal pressure and sound pressure level [20], intensity level (in dB) was also computed. Feature extraction was performed with Praat [21], identical to that in [11], except for the fact the same (default) voicing threshold was applied to f_o-dependent (L_1L_2) and and f_o-independent (intensity level, CPPS) features.

Voiced intervals of at least 20 ms were identified based on the output of Praat's pitch tracking. Next, the last voiced interval in the final second of each talkspurt was found, the CPPS, L_1L_2 and intensity level trajectories were interpolated with cubic splines (without smoothing) and sampled at 15 evenly spaced points. Frames that included pitch-halving errors were removed. If this procedure resulted in a loss of more than 10% of frames in a voiced interval, the entire interval was excluded.

Only intervals longer than 78 ms (corresponding to the minimum of 15 50-ms frames with a 2-ms step) in which all the feature values were within three standard deviations of speaker's mean were included in the final analysis. Additionally, data from one Swedish speaker were excluded due to evidence of strong pitch halving. The final data set consisted of: 234 BSS and 649 WSS intervals in Swedish, and 384 BSS and 618 WSS intervals in German.

2.3. Modeling

The data were modeled using Generalized Additive Mixed Models (GAMM, [22]) using the *mgcv* package in R [23]. Separate models were fitted for each response variable (intensity level, CPPS, L_1L_2). The following model structure was used, using the `mgcv::formula.gam` syntax (`s`: smooth term, `te`: tensor product smooth, `bs='fs'`: factor smooth):

```
Y ~ s(time) +
  s(time, by=is.swe) +
  s(time, by=is.wss) +
  s(time, by=is.swe.wss) +
  te(time, voiced.dur) +
  s(time, speaker, bs='fs', m=1) +
  s(time, speaker, by=is.wss.ord,
    bs='fs', m=1)
```

The models included effects for language (`is.swe`) and interval type (`is.wss`) as well as for the difference between the BSS/WSS contrast across the two languages (`is.swe.wss`). Since we have no specific hypotheses about contour shapes as opposed to the overall level, we use binary difference smooths [24]. They combine both of these differences and do not result in inflated type-I error rates associated with separate tests for the parametric and the non-linear components [25]. In addition, the models included a tensor product interaction between time and the duration of the voiced interval (`voiced.dur`, in $log_2(\text{s})$), as well as random smooths for speakers (`speaker`), grouped by interval type (`is.wss.ord`, modelled using a reference and random difference smooths, following the recommendation in [25]).

All the models were fitted with the `mgcv::bam` function, using fast restricted maximum likelihood estimation (fREML) and discretization of covariate values (i.e. with the `discrete` parameter set to true). A scaled t model (`family = 'scat'`) was used to deal with non-linearity of the residuals. Autocorrelation in the data was controlled for with a

first-order auto-regressive model (AR1), with ρ set to autocorrelation at lag of one in a model without the AR1 component.

Secondary data and the scripts used in the analysis are available at: https://doi.org/10.5281/zenodo.7870015.

3. RESULTS

The predicted curves for intensity level, CPPS and L_1L_2 are shown in Figure 1.

The observed intensity effects reflect the expected patterns associated with syllabic organisation in speech: intensity level increases at the beginning of the voiced interval and falls again towards its end. Indeed, median duration of the voiced interval was equal to 0.2 seconds, which corresponds roughly to average syllable duration in speech. The end point is somewhat lower than the start point, most likely corresponding to energy declination across prosodic phrases. The difference between German and Swedish was significant for the reference BSS category ($p < 0.001$). Visual inspection of Figure 1 as well as of the difference smooth indicates that the curve for Swedish is somewhat lower (by about 2 dB) and shows a more gradual decrease of intensity throughout its duration, as opposed to the German curve, which decreases more sharply towards its end. Neither the BSS/WSS contrast in German ($p = 0.09$) nor the difference in marking of the contrast across the languages ($p = 0.6$) were significant.

CPPS mirrors the inverted-U pattern present in the intensity curves, reflecting the fact that centers of syllabic nuclei are characterised by more periodic phonation than the peripheries. We found no evidence for a difference in the BSS category across the languages ($p = 0.7$). The BSS and WSS intervals differed significantly for German ($p < 0.001$): within-speaker silences were characterised by higher CPPS values in the middle portion of the voiced interval. However, the difference curves reveal that this difference is very small (less than 1 dB). There was no evidence that the BSS/WSS contrast is realised differently in Swedish ($p = 0.4$).

L_1L_2 shows the opposite pattern to intensity level and CPPS with low values in the middle of the voiced interval and high values towards the edges. This is again expected given that a strong fundamental (and, consequently, high L_1L_2) is indicative of more breathy phonation. The language effect for the BSS category was significant ($p < 0.001$): the curve for Swedish is somewhat higher and is characterised by a less steep increase towards the end. The BSS/WSS contrast in German was also significant ($p = 0.02$), with lower values for WSS intervals but, similar to CPPS, the difference was smaller than 1 dB on average. We found no evidence for language-specific difference with respect to the BSS/WSS effect ($p = 0.7$).

4. DISCUSSION

This paper is part of project where we are exploring voice quality as a dimension of prosodic expression, following the voice prosody hypothesis [4, 5, 6]. Our approach has been to examine short-term variations in acoustic voice quality features as correlates of various prosodic functions (e.g. turn-yielding, turn-holding, and word and utterance level prominences). To avoid the difficulties in estimating voice quality features from running speech, we have based our features on signals from miniature accelerometers attached to the speaker's neck below the level of the glottis where the influence of the vocal tract is minimal.

The results presented in this work indicate that voice quality varies over syllable-sized units roughly in the same way as intensity level. The middle part of the last voiced interval was associated with higher intensity level and, consequently, more periodic phonation, indicated by higher CPPS and lower L_1L_2. This effect is present in both languages and regardless of the different segmentation methods employed in the two data sets.

The other effects investigated here, although present, are very small and did not vary across languages. In line with previous research [8, 9, 11], between-speaker intervals were characterised by less periodic phonation (lower CPPS, higher L_1L_2) but the differences were below 1 dB, on average. However, it is possible that the feature extraction methods employed, and in particular the removal of frames with pitch-halving, reduced the influence of creak on the results. Thus, the role of creak as a turn-taking cue might be greater than indicated by our results. It is also conceivable that the the BSS/WSS contrast is realised differently as a function of the duration of the last voiced interval, or depends on the duration of the talkspurts, with greater aperiodicity being more likely towards the end of longer turns [26]. Finally, voice quality cues might be employed over larger temporal domains than the last voiced interval of a talkspurt. We leave these questions open for future research.

Of all the features investigated here, intensity level showed the least amount of separation across the between and within-speaker categories. This stands in stark contrast to the results in [11],

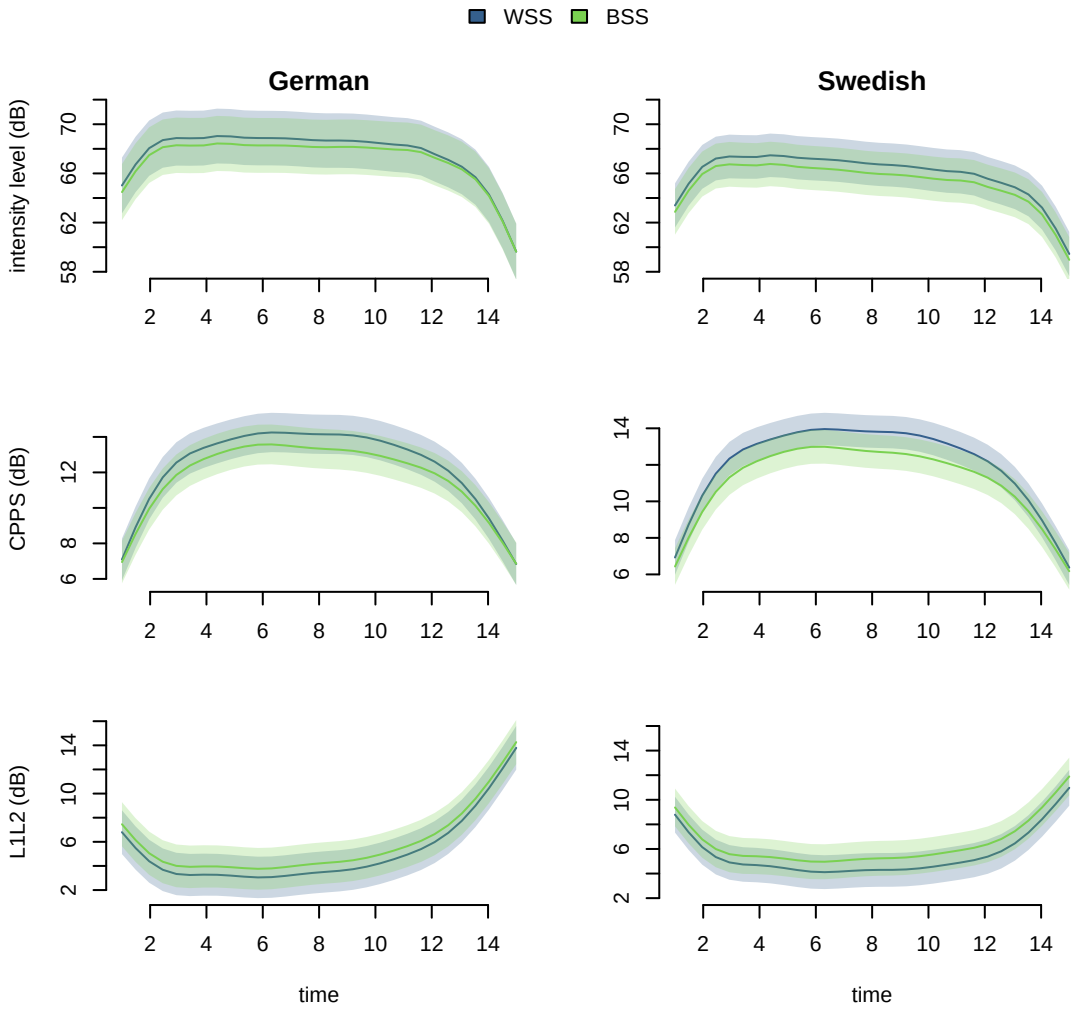

Figure 1: Model predictions for intensity level (top), CPPS (middle) and L_1L_2 in Swedish (left) and German (right) for between-speaker and within-speaker silences (BSS and WSS, respectively).

where intensity made the biggest contribution to discriminating between turn-taking categories. This could be partly explained by the fact that intensity level (as well as other f_o-independent features) in [11] included values extracted from intervals detected with a lower voicing threshold in order to capture even less periodic phonation types (this procedure was not used in the present study to produce time-aligned trajectories for f_o-dependent and f_o-independent features). Such intervals, which involve very breathy or creaky voice quality, are likely to be much quieter than the voiced intervals investigated here. Furthermore, the analysis in [11] did not control for covariates such as interval duration and did not model the grouping structure in the data.

The fact that the CPPS and L_1L_2 curves in Figure 1 exhibit greater separation across the BSS/WSS classes might suggest that voice quality does in fact carry non-redundant information related to turn-taking. Similarly, while the shape of the L_1L_2 contour is essentially a mirror image of the intensity level curve, the CPPS shows a more distinct, symmetrical pattern, possibly indicating that CPPS is to a less extent determined by sound pressure level (and subglottal pressure) variation. Nonetheless, given the small effect sizes, the effects are most likely of little practical importance in speech communication.

5. ACKNOWLEDGEMENTS

This work was funded by Swedish Research Council project *Prosodic functions of voice quality dynamics* (VR 2019-02932) to Marcin Włodarczak.

6. REFERENCES

[1] J. Local, "Variable domains and variable relevance: Interpreting phonetic exponents," *Journal of Phonetics*, vol. 31, no. 3, pp. 321–339, 2003.

[2] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica*, vol. 62, no. 2–4, pp. 215–226, 2005.

[3] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, 2011.

[4] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2012, pp. 378–423.

[5] A. N. Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The voice prominence hypothesis: The interplay of f0 and voice source features in accentuation." in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 3527–3531.

[6] A. N. Chasaide, I. Yanushevskaya, and C. Gobl, "Prosody of voice: Declination, sentence mode and interaction with prominence," in *Proceedings of the XVIIIth International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK, 2015.

[7] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[8] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[9] M. Heldner, M. Włodarczak, Š. Beňuš, and A. Gravano, "Voice quality as a turn-taking cue," in *Proceedings of Interspeech 2019*, Graz, Austria, 2019, pp. 4165–4169.

[10] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," in *Proc. Interspeech 2017*, 2017, pp. 2351–2355.

[11] M. Włodarczak and M. Heldner, "Contribution of voice quality to prediction of turn-taking events," in *Proceedings of Speech Prosody 2022*, Lisbon, Portugal, 2022, pp. 485–489.

[12] A. S. Fryd, J. H. Van Stan, R. E. Hillman, and D. D. Mehta, "Estimating subglottal pressure from neck-surface acceleration during normal voice production," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. 1335–1345, 2016.

[13] V. S. McKenna, A. F. Llico, D. D. Mehta, J. S. Perkell, and C. E. Stepp, "Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 12, pp. 3404–3416, 2017.

[14] D. D. Mehta, V. M. Espinoza, J. H. Van Stan, M. Zañartu, and R. E. Hillman, "The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. EL386–EL392, 2019.

[15] M. Borsky, M. Cocude, D. D. Mehta, M. Zañartu, and J. Gudnason, "Classification of voice modes using neck-surface accelerometer data," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5060–5064.

[16] M. Włodarczak, B. Ludusan, J. Sundberg, and M. Heldner, "Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound," *Journal of Voice*, In press.

[17] K. Laskowski, "Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2011.

[18] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011, pp. 2837–2840.

[19] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, 1996.

[20] J. Sundberg, "Flow glottogram and subglottal pressure relationship in singers and untrained voices," *Journal of Voice*, vol. 32, no. 1, pp. 23–31, 2018.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Computer program, 2022.

[22] S. N. Wood, *Generalized Additive Models. An Introduction with R*. Boca Raton: CRC Press, 2017.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[24] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[25] M. Sóskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, pp. 1–22, 2021.

[26] K. Aare, P. Lippus, M. Włodarczak, and M. Heldner, "Creak in the respiratory cycle," in *Proceedings of Interspeech 2018*, 2018, pp. 1408–1412.