FINAL LENGTHENING IN LINE END PERCEPTION OF RE-SYNTHESIZED RECITATIONS OF GERMAN POEMS

Nadja Schauffler[1], Julia Koch[1], Nora Ketschik[1], Toni Bernhart[2], Felix Dieterle[3], Gunilla Eschenbach[3], Anna Kinder[3], Sandra Richter[3], Gabriel Viehhauser[2], Thang Vu[1], Jonas Kuhn[1]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Institute of Literary Studies/Digital Humanities, University of Stuttgart, Germany
[3]German Literature Archive (DLA), Marbach, Germany
nadja.schauffler@ims.uni-stuttgart.de

ABSTRACT

This paper presents a perception study investigating the perceptual relevance of lengthening at the end of lines with enjambment in poem recitations. An enjambment occurs when the end of a line disrupts a syntactic unit. In recitation, a speaker has different possibilities to realize this conflict between syntactic coherence on the one hand, and versification on the other. In a previous study, lengthening of the verse-final segment was found in lines both with and without enjambment. In the present study, we investigated whether this lengthening effect is perceptually meaningful in the prediction of line breaks. We re-synthesized the original recitation and manipulated the verse-final segments' duration. While predicting line breaks seemed to be an overall difficult task for listeners, we found that verse-final lengthening increases the success rate and was thus perceived by listeners as a prosodic cue to versification, even when it occurred within syntactic constituents.

Keywords: perception, speech synthesis, enjambment, poetry, digital humanities

1. INTRODUCTION

Lyric poetry is characterized by the use of lines and stanzas as primary structural units, creating an additional layer of structure and aesthetics beyond the syntax of clauses and sentences. This can lead to tensions between the linguistic form (syntax) and the poetic form (versification). Enjambments - when the syntactic unit continues past the end of a line - are an example of this tension. This mismatch between verse and syntax offers the speaker the opportunity to emphasize either the syntactic unity or the verse structure.

It has been suggested that there is a third way for speakers to serve both layers by simultaneously using prosodic cues signalling continuation as well as prosodic cues signalling discontinuation [1, 2]. In a corpus study investigating one professional speaker and his recitations of poems by the German poet Friedrich Hölderlin, Schauffler et al. [3] looked at prosodic features used by the speaker to mark the end of the line. They compared lines with enjambment to lines without enjambment and analyzed cues typically found at prosodic boundaries, namely lengthening of the phrase-final segment(s) (cf. [4, 5, 6]), the insertion of a silent pause (e.g., [7]) and F_0 reset (e.g., [8, 9]). They found that the speaker indeed reliably marks the end of a line by means of final lengthening, even in cases of enjambment, while on the other hand, silent pauses and F_0-register effects were used significantly less often in lines with enjambment.

In the study at hand, we are investigating whether verse-final lengthening is perceived by listeners as a cue marking the end of the line. It has been shown before that phrase-final lengthening is a cue in syntactically ambiguous structures, such as in the perception of bracketed lists [10]. In an online perception experiment, we have now tested whether verse-final lengthening also has a structuring effect when it occurs within syntactic units, such as in enjambments. To this end, we used the original recordings from [3] and re-synthesized them by cloning the speakers' F_0, energy and duration for each segment [11]. In a second step, we manipulated the duration of the verse-final segments (nucleus and coda) in three steps in order to investigate whether the segments' duration affects the identification of line ends.

This experiment also serves as a baseline study to explore human-in-the-loop re-synthesis as a methodological tool in hypothesis testing, as suggested in the »textklang« project.[1] Our synthesis approach allows to generate test items for perception experiments in which particular prosodic parameters can be manipulated while keeping the prosodic realisation of the original. This can be especially advantageous for questions in the realms of digital humanities where natural data is often too limited to systematically evaluate hypotheses.

2. STIMULI

The stimuli were taken from the same corpus of Hölderlin poems as in the study by [3]. Only lines with free verse and no rhymes were considered.[2] In order to avoid traces of other prosodic phrasing cues, we picked two-line pairs with enjambment where neither a silent pause nor F_0 reset was used by the speaker (cf. [3]). Furthermore, we excluded lines in which the speaker made an audible caesura in close proximity to the line break (within two words before or after the line end). After re-synthesis (see below) the stimuli were auditorily evaluated. We excluded cases where the synthesis model produced an audible pause at the end, and one case where the re-synthesized verse-final word was apparently unknown (and thus lead to an unnatural re-synthesis). This resulted in 16 items. We divided them into two groups, 11 items with "strong" enjambments, i.e. the line-break disrupts a syntactic constituent which cannot be further decomposed, and 5 items with "weak" enjambments, where the line break takes place between syntactic phrases. See 1[3] for an example of a strong enjambment and 2[4] for an example of a weak enjambment.

(1) Den Stromgeist fern, und schaudern regt [im Nabel [der Erde]$_{NP}$]$_{PP}$ der Geist sich wieder.

(2) Doch [die ewige Sonne]$_{NP}$ goß
[ihr verjüngendes Licht]$_{NP}$ über das alternde

3. RE-SYNTHESIS AND MANIPULATION

3.1. Synthesis approach

We used the open source code and pretrained multilingual models provided by the IMS Toucan Speech Synthesis Toolkit release v2.2 [12] to create our stimuli. This toolkit implements a text-to-speech (TTS) pipeline consisting of multiple components: First, the input text is transcribed to a phonetic transcription according to IPA. Then a FastSpeech 2 based TTS model [13] generates spectrograms from the input phonemes and finally a HiFi-GAN vocoder [14] converts the spectrograms to waveform. A main advantage of FastSpeech 2 is its inherently high level of controllability due to dedicated submodules for predicting duration, F_0 and energy. The implementation in IMS Toucan comes with a lightweight reconstruction-based aligner that is trained jointly with the TTS module to extract phone durations from the reference audio [12, 15]. Thereby, alignment is directly integrated into the TTS pipeline instead of relying on an external tool for forced-alignment as seen in the original FastSpeech 2 paper [13]. To exactly re-synthesize a reference recording, we follow the approach on prosody cloning described in [15]: Duration, F_0 and energy values are extracted from the original recording. To obtain phone-wise controllability instead of dealing with individual spectrogram frames, pitch and energy values are averaged over the spectrogram frames that belong to a single phoneme according to the aligner as first proposed in [16]. These extracted prosodic values are then used to overwrite model predictions given by the duration, pitch and energy predictors respectively. Koch et al. [11] further explore the idea of overwriting prosodic values in a human-in-the-loop setup. They first clone the prosody of a reference audio and in a second step, a human expert can adjust prosodic values of individual phones using the same overwriting technique in order to make fine-grained manipulations to the prosody of the original recording. We followed this approach in our experiment.

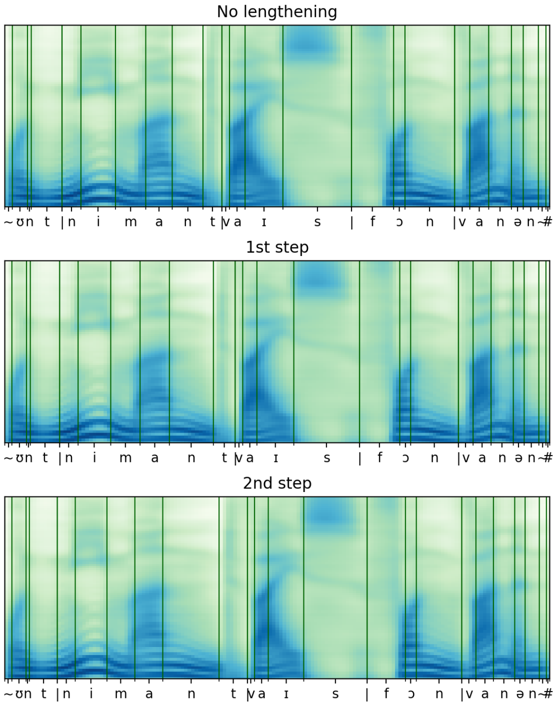

Figure 1: Spectrograms of the three different manipulations of the word *niemand* in the sentence "[...] und niemand weiß von wannen [...]"[5]. All phones in the last syllable rhyme of this word (a, n and t) are affected by the manipulations. The green bars mark phone boundaries. | denotes word boundaries.

3.2. Manipulation

We manipulated the final syllable rhyme in each line of the re-synthesized first lines since this is where final lengthening has been shown to be particularly pronounced (e.g. [4, 6, 17]). Each segment in this rhyme was manipulated in three steps. Each step is based on the original data:

no lengthening The original duration of the line-final segments was replaced by the mean duration of the respective segment in the respective syllable position (syllable-final for segments in the coda, syllable-vowel for non-final vowels) as taken from the original corpus.

1st step The segments were lengthened to the mean duration of the respective segments at the end of a line, i.e. to the mean z-score of these segments compared to non-final segments.

2nd step The segments were further lengthened corresponding to the upper quartile of the z-score of the respective final segments' duration in the original data.

Segment duration of the original data was extracted by means of the Festival [18] version of the University of Stuttgart [19]. Since our TTS models segment duration in terms of number of spectrogram frames we convert the extracted durations to spectrogram frames with the python library librosa [20] time_to_frames function according to equation 1 with $frames_{phone}$ denoting the number of frames assigned to a given phone and dur_{phone} the duration of that phone in seconds as extracted with Festival, sr being the sampling rate and hop_length meaning the number of samples between successive frames. With our settings of $sr = 16kHz$ and a $hop\_length = 256$ one spectrogram frame depicts 0.016 seconds, and hence, phone durations are modelled accurately to 16 milliseconds in our approach. To improve the audio quality of the synthetic stimuli, we perform super resolution to $48kHz$ at the stage of vocoding as shown in [21, 12].

$$(1) \quad frames_{phone} = floor(dur_{phone} * sr/hop\_length)$$

Figure 1 shows a comparison of spectrograms visualizing the different steps of manipulations. In this example the line break occurs immediately after the word *niemand* (no one). Hence, we manipulate all phones within the final syllable rhyme of this word, i.e. the phones a, n and t. While the difference for the phone a is rather small we observe a clearly visible difference in duration for n and t.

4. PARTICIPANTS AND PROCEDURE

For our online perception study, participants were recruited via several mailing lists and directed to www.soscisurvey.de where the experiment was made available. The call for participation was addressed to German L1 speakers. In a questionnaire following the experiment 40 participants indicated that they have a background in literary studies, and 32 participants indicated a linguistic background. The items with their three manipulations were distributed over three lists in a between-subject Latin Square design so that each participant only listened to one manipulation per item and to 16 items in total. The items were presented in randomized order in an online questionnaire which was implemented using SoSci Survey [22]. 112 participants took part in the experiment from beginning to end. They were randomly assigned to one of the three lists. The participants were instructed to listen to the audio. They were given the two verses in written form in one line without the line break. If more context was needed in order to make sense of the two lines parts of the preceding or following line (or both) were given in brackets. A sequence of four words was underlined and participants were instructed to choose the one they thought was the last in the first verse. For each item they could indicate how certain they are with their choice using a slider on a 0-10 scale. They could listen to the audio files as often as necessary. After the experiment, participants were asked about how difficult the task was, and about their background (see above). They were also given the possibility to comment on the experiment.

5. ANALYSIS

Four participants indicated in the comment section that they did not listen to the audio file but judged the line break according to their own reading. These participants were excluded from the analysis so that the answers of 110 participants were statistically analysed, namely 585 cases for the baseline condition without lengthening, 585 cases for the 1st step lengthening and 590 cases for the 2nd step lengthening.

The statistical analysis was performed in R 4.2.2 [23], using the function glmer from package lme4 [24]. In order to investigate the relationship between the identification of the line break and the amount of verse-final lengthening, we performed a generalized linear mixed effects analysis using the logit link function with whether the last word in the verse was correctly identified or not as the dependent variable. As fixed factors we included *lengthening* (no lengthening, 1st step and 2nd step), and the *type of enjambment* (strong or weak). Since it may be possible that cues correlating with final lengthening may still be present in the signal, we also included the last segments' *duration in the original* as phoneme-specific z-score as fixed factor. As random factors we included intercepts for *participants* and *items*. All factors were tested for their significance by means of likelihood ratio tests by comparing the model including the effect in question to the model without it.

6. RESULTS

Most participants thought the task was rather difficult or depending on the item sometimes rather difficult and sometimes rather easy (mean 3.6 on the 1-5 point scale where 1 is very easy and 5 is very difficult).

Figure 2 gives an overview of the effect of the amount of lengthening (x-axis) on the probability of correctly recognizing the last word of the line (y-axis). We can see that the text itself hardly gives any indication as to where the line break occurs since when there was no lengthening (intercept: β=-0.57, SE=0.79, p=.47) participants correctly identified the last word in only 40% of the time. The ability to identify the last word increases significantly with the first step length manipulation to about 52% (β=0.63, SE=0.15, p<.0001) and increases further with the second step length manipulation to about 59% (β=1.25, SE=0.16, p<.0001). There was no effect for type of enjambment. Also the original duration of the line-final segments did not affect the probability to correctly identify the last word.

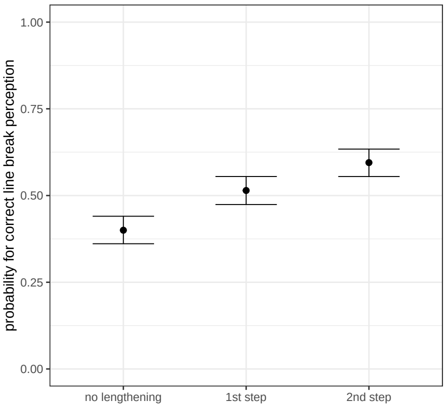

Figure 2: The probability for the indication of the correct line end by condition.

7. DISCUSSION AND CONCLUSION

Given the complexity of the poetic language, the lack of meter, and the limited context the task of finding the correct line break was rather difficult as reflected in the degree of difficulty reported by the participants on the one hand and the overall low success rate on the other hand. In order to make the task more approachable, only four options were given per line. These options may, of course, also have an effect on the probability for choosing the correct word. The other words were not manipulated in any way other than being re-synthesized so that we cannot exclude the presence of phrase-final cues in these words. Since the non-final words were always the same, differences between the length conditions can, however, be attributed to the manipulation of the line-final segments' duration. This manipulation of length had a significant effect: it facilitated the identification of the line end. The more pronounced the last segments' lengthening was the more likely listeners perceived the lengthening as a prosodic cue marking the end of the line. This result suggests that final lengthening is available to listeners as a structure-giving cue in this genre-specific context, even if it is in conflict with the syntactic coherence continuing over the line break. The finding that enjambments without a silent pause can still be prosodically conveyed by using more subtle cues such as final lengthening supports the idea of a "rhythmical performance" as suggested by [2, 1].

Surprisingly, the type of enjambment (strong or weak) did not have an effect on the line end perception even though we would have expected that weak enjambments are more likely to be identified given the weaker syntactic cohesion. With only 5 items in the weak-enjambment group, the data was, however, not balanced in this respect. Further research is needed to investigate the effects of different types of enjambments on both the production of line-final prosodic cues and their effect in perception.

With respect to re-synthesis we can say that it presents a valuable tool in experimental research by making it possible to manipulate fine-phonetic detail while cloning everything else from the original. A few participants reported that the audio sounded to be of not the best quality, or "tinny", and only one participant mentioned that some words sounded manipulated. In our next step we will work on improving the synthesis model with respect to speech quality and accuracy of prosody cloning.

8. ACKNOWLEDGEMENTS

This research is supported by funding from BMBF for the »textklang« project.

9. REFERENCES

[1] R. Tsur and C. Gafni, "Enjambment – irony, wit, emotion. A case study suggesting wider principles," *Studia Metrica et Poetica*, vol. 5, pp. 7–28, 01 2019.

[2] R. Tsur, *Poetic Rhythm: Structure and Performance – An Empirical Study in Cognitive Poetics*. Brighton and Portland: Sussex Academic Press, 2012.

[3] N. Schauffler, F. Schubö, T. Bernhart, G. Eschenbach, J. Koch, S. Richter, G. Viehhauser, T. Vu, L. Wesemann, and J. Kuhn, "Prosodic realisation of enjambment in recitations of German poetry," in *Proc. Speech Prosody*, Lissabon (Portugal), 2022.

[4] D. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1220, 1976.

[5] A. Turk and S. Shattuk-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, pp. 445–472, 2007.

[6] K. J. Kohler, "Prosodic boundary signals in German," *Phonetica*, vol. 40, pp. 445–472, 1983.

[7] M. Nespor and I. Vogel, *Prosodic Phonology*. Dordrecht, Holland: Foris, 1986.

[8] C. G. Berg, Rob van den and T. Rietveld, "Downstep in Dutch: Implications for a model," in *Papers in laboratory phonology II: Gesture, segment, prosody*, G. J. Docherty and D. R. Ladd, Eds. Cambridge: Cambridge University Press, 1992, pp. 335–367.

[9] H. Truckenbrodt, "The syntax phonology interface," in *The Cambridge handbook of phonology*, P. de Lacy, Ed. Cambridge: Cambridge University Press, 2007, pp. 435–456.

[10] C. Petrone, H. Truckenbrodt, C. Wellmann, J. Holzgrefe-Lang, I. Wartenburger, and B. Höhle, "Prosodic boundary cues in german: Evidence from the production and perception of bracketed lists," *Journal of Phonetics*, vol. 61, pp. 71–92, 2017.

[11] J. Koch, F. Lux, N. Schauffler, T. Bernhart, F. Dieterle, J. Kuhn, S. Richter, G. Viehhauser, and N. T. Vu, "PoeticTTS - Controllable Poetry Reading for Literary Studies," in *Proc. Interspeech 2022*, 2022, pp. 1223–1227.

[12] F. Lux, J. Koch, and N. T. Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2022.

[13] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *ICLR*, 2020.

[14] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, vol. 33, 2020.

[15] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *Proc. IEEE SLT*, 2022.

[16] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP*. IEEE, 2021, pp. 6588–6592.

[17] A. Schweitzer, *Production and Perception of Prosodic Events – Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart, 2010.

[18] A. W. Black, "The Festival speech synthesis system," www.cstr.ed.ac.uk/projects/festival.html, November 1997.

[19] Festival, "Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. IMS German Festival home page," www.ims.uni-stuttgart.de/phonetik/synthesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2010.

[20] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, L. Nickel, P. Friesch, M. Vollrath, and T. Kim, "librosa/librosa: 0.9.2," Jun. 2022.

[21] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li *et al.*, "DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021," *Proc. Blizzard Challenge Workshop*, vol. 2021, 2021.

[22] D. J. Leiner, "Sosci survey (version 3.4.04)," 2021. [Online]. Available: https://www.soscisurvey.de

[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] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear mixed-effects models using S4 classes*, 2013, r package version 0.999999-2. [Online]. Available: http://CRAN.R-project.org/package=lme4

[1] http://hdl.handle.net/11022/1007-0000-0007-F6C5-5

[2] In one item, we manually exchanged the rhyming word with a phonetically similar non-rhyming word.

[3] *The stream spirit far away, and shuddering in the // Navel of the earth the spirit stirs again* (own translation).

[4] *But the eternal sun poured // its rejuvenating light over the aging* (own translation).

[5] *and no one know from where [...])*