On the relationship between pointing gestures and speech production in German counting out rhymes: Evidence from motion capture data and speech acoustics

Conference Paper · October 2016

CITATIONS
0

READS
43

2 authors:

Susanne Fuchs
Centre for General Linguistics
112 PUBLICATIONS **530** CITATIONS
SEE PROFILE

Uwe Reichel
Hungarian Academy of Sciences
62 PUBLICATIONS **137** CITATIONS
SEE PROFILE

On the relationship between pointing gestures and speech production in German counting out rhymes: Evidence from motion capture data and speech acoustics

Susanne Fuchs [1], *Uwe D. Reichel*[2]

[1] Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany
[2] Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
fuchs@zas.gwz-berlin.de, uwe.reichel@nytud.mta.hu

Abstract

We investigated the interplay between pointing gestures and speech by means of motion capture and acoustics. Counting out rhymes served as a testbed, since they involve clear index finger turning points. The distance between the participant and an interlocutor (a teddy) was varied between close and far. Additionally, speaking with a normal speech rate in comparison to a fast rate was examined. Results of 1352 pointing gestures provide evidence that: a) the number of syllables realized per stroke are in general relatively stable across condition, but differences occur among subjects, b) turning points occur frequently in vowels, but also in consonants when syllables have a complex phonological structure, c) fast speech rate not only affects speech, but also leads to a shortening in pointing gesture duration, d) rhythmicity of the strokes is reduced with high speech rate and e) the impact of stroke rate on the acoustic energy contour is larger in normal than in fast speech. Distance showed no strong effects. We believe that counting out rhymes show great potential for further research in which further insight could be gained into the rhythmic and prosodic characteristics of a language as well as the coordination between pointing gestures and speech.

Index terms: pointing gestures, counting out rhymes, motion capture, syllable structure, rhythm, DCT

1. Introduction

This study investigates the relationship between manual pointing gestures of the index finger and speech acoustics in counting out rhymes. Counting out rhymes are an interesting area of investigation, because they naturally include pointing gestures with clear turning points and they can be recorded without constraining the motion of a speaker and therefore have a high external validity in comparison to very controlled, unnatural situations. Moreover, they are considered to be part of the oral poetry tradition found in many languages and cultures [7]. Kelly and Rubin [8] stated that rhythmic structure of poetry can provide evidence regarding the speech rhythm of a language. Finally, counting out rhymes are perfect games to investigate the acquisition of prosodic rules, since the rhymes are frequently used in childhood. The intention of the game is to select one person pseudo-randomly out of two (or a whole group). Since in counting out rhymes the number of syllables and words for each phrase often changes (see Table 1), it is to some extent unpredictable where the whole rhyme will end and which person will be "out". The person who counts speaks the rhyme and moves his/her index finger back and forth between the other and him/herself with very clear turning points. It is therefore an ideal testbed to investigate the relationship between speech production and pointing gestures. In particular it allows investigating the following questions:

(1) Does the number of turning points resemble the number of syllables? Is this behavior speaker specific or rather stable among speakers and conditions?

(2) Where do turning points occur within the speech flow (only in the vowel or also elsewhere)?

(3) To what extent is the relationship between speech and pointing gestures affected by time pressure and distance between two players?

These questions are motivated by the following proposals from the literature. Rochet Cappelan and colleagues [9] proposed an optimal 2:1 frequency relation between speech (jaw) and pointing gestures. They tested this relationship in an experiment in which adult participants had to point to a target while naming it. The target consisted of either 1, 2, 3 or 4 /pa/ syllables. Based on temporal measures of the pointing gesture and the jaw motions, the authors confirmed their proposal. They write: "… two syllables might be the maximum number of syllables that could be realized on one finger pointing motion without affecting the duration of the pointing period." (p. 5). We wish to extend their work to counting out rhymes and are particularly interested in the stability of this relation across speakers and tasks (questions 1 and 3). We would always expect speech production to be much faster than pointing gestures. Although the two belong to the same body of a person, the two motor control systems have very different mass (heavier for the arm), dynamic behavior (soft tissue dynamics for the tongue, joints for the arm and fingers), and space within they can move (much larger for pointing gestures).

Moreover, we are interested in how speakers coordinate their pointing gestures with the speech flow [6,10]. Krivokapic et al. (2016) [6] hypothesized that the maximum displacement of the index finger motion would be coordinated with the vocalic gesture or the tonal target of the stressed vowel, but not the consonantal gesture. They investigated the coordination between pointing gestures and oral gestures in bisyllabic words (CVCV), with either stress on the first or second syllable. Their findings provide evidence for a stable coordination of tonal targets (measured as f0 peaks) and the maximum displacement of the pointing gesture. We wish to follow up on this, but also to examine natural speech material which also contains voiceless portions as well as syllables with complex onsets and/or codas. It is unclear whether or not a similar coordination between pointing targets and the vowel occurs in closed or complex syllables (question 2).

2. Methodology

2.1. Experimental setup

Participants stood in front of a chair with a teddy bear and were instructed to play the counting out rhyme game with the bear as a fictitious person. We proposed various counting out rhymes, but recorded only the ones the participant knew (see Table 1). These counting out rhymes were recorded in four different conditions in successive order: a) close distance & normal speech rate, b) far distance & normal speech rate, c) far distance & fast speech rate, and d) close distance & fast speech rate.

In the close distance condition, the teddy was placed about 1m in front of the participant while in the far distance it was placed ca. 2m away. The latter two conditions also involved a faster speech rate. To evoke fast rate, subjects were instructed to imagine that they want to finish quickly with the counting out rhyme and start the following game.

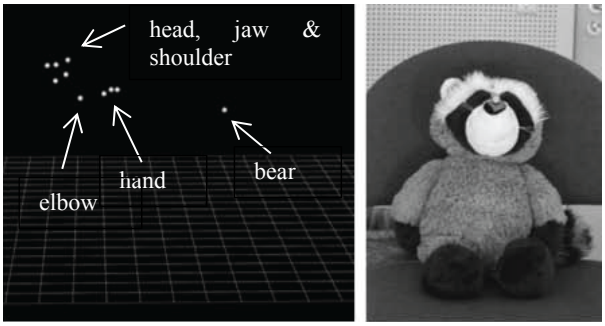

Figure 1: *Left: Display of a subject with markers located on different body parts pointing to a bear. Right: Teddy bear with a marker at the nose.*

Pointing gestures were measured by means of a motion capture system (OptiTrack, *Motive* Version 1.9.0) with 12 cameras (Prime 13). Motion data were recorded with a sampling rate of 200 Hz and a precision of 0.4 mm after calibration. One camera of the twelve was used as a video camera (200 Hz) to protocol the whole recording session. Acoustics were simultaneously recorded by means of a Sennheiser microphone. The sampling rate was 44.1 kHz.

Three markers were placed on a frontlet (one anterior, one posterior and one at the right lateral side), one marker was glued on the chin, one at the right shoulder joint (for right handers, and the left for left handed people), one at the wrist, one at the finger joint of the index finger, and one at the tip of the index finger. In the analyses we describe here, we focused only on x, y, and z motions of the index finger motion.

2.2. Participants and speech material

So far we have recorded five females with no known history of self-reported speech, language or hearing disorders. All were between 35 and 50 years old and worked in academia.

2.3. Preprocessing and annotation

Motive output files were saved in c3d format and subsequently converted into Matlab using the *Biomechanical Toolkit* [1]. The marker of the index finger was manually selected and it was checked for artefacts due to hidden movements during the relevant pointing gestures. To label the data and select the respective turning points, the 3D matrix was converted into a motion rate vector and saved as a wav-file. All data were then annotated in PRAAT (version 5.3.53 [2]). On the basis of the motion rate we labeled the velocity minima which correspond to the index finger endpoints (turning points). The interval between two turning points will hereafter be called a stroke. In a further step we added the text which was spoken within a stroke. If a turning point occurred in the middle of a word or sound, we added a full stop to the text marking the continuation of that sound to the next stroke. An example is displayed in Figure 2.

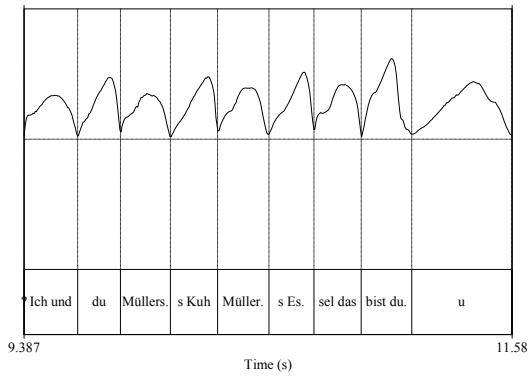

Figure 2: *Example of the motion rate vector. Vertical lines mark the labelled velocity minima. Lower track: the text of the rhyme (rhyme 1) was added for each stroke.*

Syllable nuclei were detected automatically. For this purpose the energy (root mean squared deviation: RMSD) in an analysis window was compared to the energy in a longer reference window with the same time midpoint moved along the bandpass filtered signal. If the RMSD in the analysis window was above a threshold relative to the RMSD in the reference window, a syllable nucleus was set. Further details of this procedure are provided in [3].

2.4. Location of turning point within the speech flow

A common proposal in the literature is that turning points of pointing gestures occur during vowels. Since vowels are produced with a high intensity in the acoustic envelope, we suppose that turning points would coincide with these values. For this purpose, we automatically extracted the intensity value of each turning point, subtracted it by the intensity minimum in the interval, multiplied it by 100 and divided it by the intensity range (as the difference between the max-min). The higher the output value (in percent), the closer the turning point to the intensity maximum within the stroke.

2.5. Quantification of rhythm

2.5.1. Rhythmicity: Pairwise modulo variability index

To quantify the rhythmicity of syllables and strokes we adopted the Pairwise Variability Index (PVI, [4]) a well-established measure of speech rhythm. This index measures the mean deviation of duration of neighbouring segments in an utterance. High PVI values indicate low rhythmicity and low PVI values high rhythmicity. For counting out rhymes however, this measure did not capture rhythmic variation based on varying assignment of beats per syllable or stroke. To give an example from rhyme 1, "Ich und du" consists of four

beats, “du” receiving two. Even though this results in a perfectly rhythmic four-fourth beat, PVI values would be large due to the large duration differences between the last two syllables. In order to take such a regular temporal variation into account, we extended the PVI to the Pairwise Modulo Variability index (PMI) which is normalized for speech rate (nPMI). Differences between the two indexes are illustrated in Figure 3. For a sequence X = $x_1 \ldots x_n$ (here: stroke intervals) it is calculated as follows:

$$\mathrm{nPMI}(X) = \frac{\sum_{i=1}^{n-1} [\frac{\mathrm{flip}(\max(x_i, x_{i+1}) \% \min(x_i, x_{i+1}))}{\min(x_i, x_{i+1})}]}{n-1}$$

’%’ is the modulo operator. The difference is not directly measured between the two duration values but between the larger value and the closest whole-number multiple of the smaller value. To allow this closest multiple to be larger than the greater value we had to extend the standard modulo calculation by a ’flip’ operation that replaces modulo values

$$v > \frac{\min(x_i, x_{i+1})}{2} \quad \text{by} \quad \min(x_i, x_{i+1}) - v$$

To give a concrete numerical example for this operation: the comparison of 1 with 0.4 (closest multiple 0.8) would yield a similar result as 1 with 0.6 (closest multiple 1.2). In both cases 0.2 is left.

The divisor $\min(x_i, x_{i+1})$ normalizes for speech rate, so that the nPMI is the same for pairs with the same relative durational difference, e.g. nPMI(1, 0.6) = nPMI (2, 1.2) = 0.33. As the PVI, also the nPMI gives the mean value of all pairwise differences.

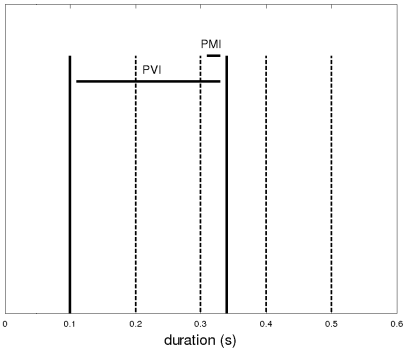

Figure 3: *Duration difference measurements PVI and PMI. PMI measures the difference between the larger value and the closest whole-number multiple of the smaller value.*

2.5.2. Influence of the strokes on the energy contour

To quantify the influence of strokes on the acoustic properties of the speech signal, we performed a digital cosine transform (DCT) on the energy contour [5]. The energy contour was taken, because it is a continuous signal and not interrupted, as e.g. fundamental frequency by voiceless events.

Then we calculated stroke influence *s* as the relative weight of the coefficients around the stroke rate *r* (± 1 Hz) within all coefficients below 10 Hz:

$$s = \frac{\sum_{c: r-1 \le f(c) \le r+1 \mathrm{Hz}} |c|}{\sum_{c: f(c) \le 10 \mathrm{Hz}} |c|}$$

Greater *s* values correspond to a greater impact of stroke rate on the energy contour.

3. Results

3.1. Relationship between number of syllables & strokes

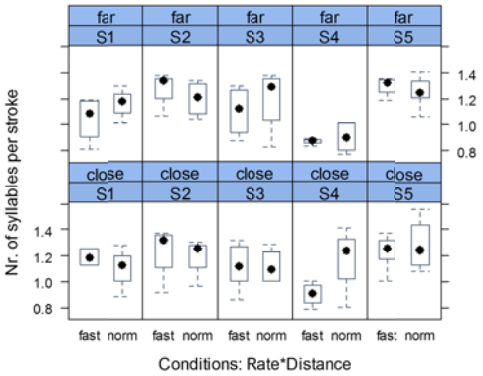

Figure 4: *Boxplots with number of syllables per stroke split by Speaker (S1-S5), Rate (fast, norm) and Distance (far, close).*

Figure 4 displays the results for how many strokes are on average spoken per pointing gesture. Except for S4, speakers are relatively stable across conditions, which speaks for a close link between the number of realized syllables and pointing gestures. The results are also in agreement with [9] that two syllables might be an upper bound for a pointing gesture.

3.2. Position of turning points within the speech flow

In our dataset 1352 pointing gestures were analyzed. In about 50% of the cases the turning points occurred between 80-100% within the intensity range, i.e. close to the intensity maximum each pointing gesture interval. We assume that these high intensity values correspond to vowels and confirm earlier proposals [6]. However, it is also evident that the other half of the data may not show turning gestures that correspond so clearly to the syllable nucleus, particularly in the fast rate. In comparison to previous studies, our speech material also involved coda consonants and complex syllable onsets. First visual inspections reveal that, particularly in closed syllables, turning points occur in coda consonants.

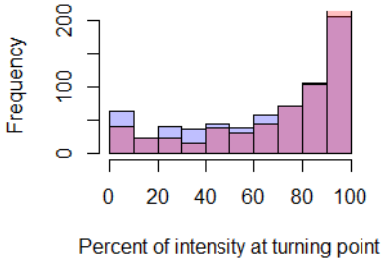

Figure 5: *Histogram for the percentage of the intensity at the turning point with respect to the overall intensity range within the stroke interval. Red: normal, Blue: fast.*

3.3. Effect of time pressure and distance

While distance shows no effect on the duration of the pointing gesture, speech rate does. Figure 6 (top) shows a shortening of the pointing gesture with faster speech. The shortening could be required if the number of syllables spoken in a stroke (Figure 4) should remain rather stable.

Furthermore, increased rate leads to a lower rhythmicity (higher nPMI values) in strokes (Figure 6 middle), and in normal speech rate to a wider distance. These effects also have consequences for speech (Figure 6 bottom). They result in a reduced impact of stroke rate on the acoustic energy contour.

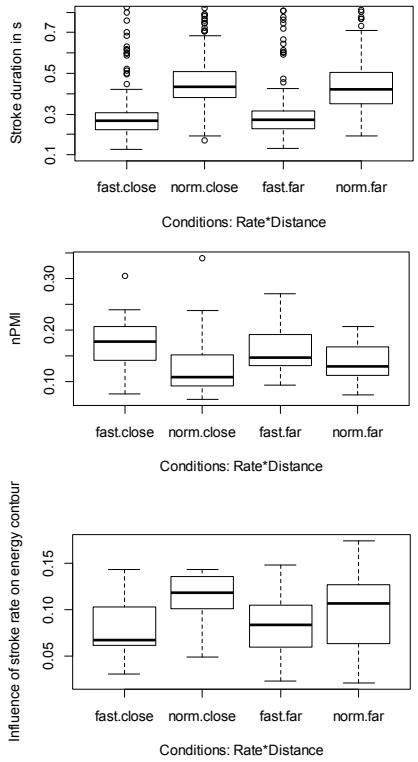

Figure 6: *top: Boxplots for stroke duration, middle: normalized pairwise modulo variability index, bottom: effect of stroke rate on energy contour split by rate and distance.*

4. Conclusion

The results of this study reveal a relatively stable, speaker-specific realization of the number of syllables per pointing gesture across conditions. When subjects speak faster, they also shorten their pointing gestures to maintain this ratio. Turning points are often produced within the vowel of a syllable, but can also occur elsewhere. Faster speech rate increases the likelihood of turning points being produced in consonants, probably because vowels are phonemes which are heavily affected by faster rate: they are reduced. Apart from the relative stability, fast speech rate reduces the rhythmicity of the pointing gestures and their impact on the energy contour of the speech signal. Such findings may be comparable to two motor systems with different properties, which adapt to each other, but can also reorganize with increased time pressure.

5. Acknowledgements

This work was supported by a grant from the BMBF to SF and by the Alexander von Humboldt society to UDR. Thanks to Jolanda Fuchs for providing her teddy bear Tom.

6. References

[1] A. Barré and S. Armand, "Biomechanical ToolKit: Open-source framework to visualize and process biomechanical data," *Computer Methods and Programs in Biomed.*, 114, 80–87, 2014.

[2] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer," Version 5.3.53, http://www.praat.org/

[3] U. Reichel, "Linking bottom-up intonation stylization to discourse structure," *Computer, Speech, and Language*, 28, pp. 1340–1365, 2014.

[4] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology 7*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 515–546.

[5] C. Heinrich and F. Schiel, "The influence of alcoholic intoxication on the short-time energy function of speech," J. Acoust. Soc. Am., vol. 135, no. 5, pp. 2942–2951, 2014.

[6] J. Krivokapic, M. Tiede, M.E. Tyrone and D. Goldenberg, "Speech and manual gesture coordination in a pointing task", In *Proc. of Speech Prosody*, Boston, paper nr. 392, 2016.

[7] P.N.A. Hanna, P. Lindner, and A. Dufter, "The meter of nursery rhymes: universal versus language-specific patterns," In *Sounds and systems: studies in structure and change*, Berlin/New York: Mouton de Gruyter, 2002, pp. 241–267.

[8] M.H. Kelly, and D. C. Rubin. "Natural rhythmic patterns in English verse: Evidence from child counting-out rhymes." *Journal of Memory and Language,* 27.6, 718–740, 1988.

[9] A. Rochet-Capellan, J.L. Schwartz, R. Laboissiere, and A. Galvan "Two CV syllables for one pointing gesture as an optimal ratio for jaw-arm coordination in a deictic task: A preliminary study." *2nd EuroCogSci07*, 2007.

[10] A. Rochet-Capellan, R. Laboissiere, A. Galvan, and J.L. Schwartz, "The speech focus position effect on jaw–finger coordination in a pointing task." *JSLHR* 51.6, 1507–1521, 2008.

7. Appendix

Table 1. *5 counting out rhymes (1st column) in German orthography. Dots: syllable boundaries within words; Line breaks: prosodic boundaries; Number of syllables = 2nd column, Number of words =3rd column, the sums and ratio between syllables and words are displayed below the text.*

Orthographic representation of counting out rhymes		**Nr. of syllables**	**Nr. of words**
(1.) Ich und du		3	3
Mül.lers Kuh		3	2
Mül.lers E.sel		4	2
der bist du		3	3
Ratio =1.3	Sum	13	10
(2.) E.ne me.ne Mis.te		6	3
es rap.pelt in der Kis.te		7	5
e.ne me.ne Meck		5	3
und du bist weg		4	4
Ratio =1.47	Sum	22	15
(3.) Ei.ne klei.ne Dick.ma.dam		7	3
fuhr mal mit der Ei.sen.bahn		7	5
Ei.sen.bahn die krach.te		6	3
Dick.ma.dam die lach.te		6	3
eins, zwei, drei		3	3
und du bist frei		4	4
Ratio = 1.57	Sum	33	21
(4.) Ei.ne klei.ne Mic.ky.maus		7	3
Zog sich mal die Ho.sen aus		7	6
Zog sie wie.der an		5	4
Und du bist dran		4	4
Ratio = 1.35	Sum	23	17
(5.) E.ne, me.ne Mo.pel		6	3
Wer frisst Po.pel		4	3
Sau.er, süß und saf.tig		6	4
Für ne Mark und acht.zig		6	5
Für ne Mark und zehn		5	5
Und du musst gehen		4	4
Ratio = 1.29	Sum	31	24