An ultrasound study of frequency and co-articulation

Motoki Saito, Fabian Tomaschek, Ching-Chu Sun, R. Harald Baayen

Eberhard Karls Universität Tübingen

motoki.saito@uni-tuebingen.de, fabian.tomaschek@uni-tuebingen.de,
ching-chu.sun@uni-tuebingen.de, harald.baayen@uni-tuebingen.de

Abstract

Anticipatory coarticulation has been reported to be affected by word form frequency. However, it remains unclear whether frequency effect also modulates carry-over (perseverative) coarticulation. To investigate the interaction of word form frequency effect and carry-over/anticipatory coarticulations, ultrasound imaging was performed on the articulation of the vowel [aː] in German verbs. Effects of coarticulation were induced by controlling the verb's suffixes and preceding pronouns. Contrary to the standard tongue contour analysis, we analyzed whole ultrasound images using Generalized Additive Models. We found more fronted tongue root, lower tongue body, and higher tongue tip in low-frequency words. By contrast, high-frequency words showed a more rounded tongue shape. This was reflected by the middle part of the tongue to be higher and the tongue root more retracted in high frequency words in comparison to low frequency words. These findings indicate more optimized tongue movements for higher frequency words.

Keywords: coarticulation, ultrasound, frequency, generalized additive mixed models

1. Introduction

According to classical models of speech production such as WEAVER++ (Roelofs 1997) or the spreading-activation model (Dell 1986), morphologically complex words are built from discrete sub-word units. This means that no representation of the whole word plays a role. Accordingly, WEAVER++ predicts that a lexical property such as whole word frequency should not co-determine phonetic realizations of a word (see also Levelt and Wheeldon 1994; Levelt, Roelofs, and Meyer 1999). Contrary to this prediction, recent studies have reported word form frequency effect on phonetic realizations (Gahl 2008; Pluymaekers, Ernestus, and Baayen 2005; Tomaschek, Wieling, et al. 2013; Tomaschek, Arnold, et al. 2018).

More specifically, Tomaschek, Tucker, et al. (2018) investigated the articulation of the stem vowel [aː] in German verbs. They found that word form frequency modulated the u-shaped articulatory trajectory of the vowel. These modulations reflected two articulatory constraints: one constraint favoring smooth trajectories through anticipatory coarticulation (Sosnik et al. 2004) and one favoring clear articulation by realizing lower minima (Lindblom 1990). The predominant pattern in low-frequency words was the constraint of clarity. In medium-frequency words, the smoothness constraint led to a raising of the trajectory. In high-frequency words, both constraints were observed, reflected by low minima and stronger coarticulation.

The study by Tomaschek, Tucker, et al. (2018) controlled for the phonetic context following the stem vowel and its effects of anticipatory coarticulation. However, the phonetic context preceding the stem vowel was not controlled for, thus neglecting carry-over effects (Öhman 1966; Song et al. 2013).

The present study closes this gap. It aims at examining word form frequency effect on coarticulation patterns, taking into anticipatory and carry-over coarticulations.

2. Methods

2.1. Recording

The material for the present study consisted of 126 German verbs with the stem vowel [aː]. To control for coarticulation effects, verbs were articulated in two pronoun conditions ('sie' [zi], 'ihr' [iːɐ]) and two suffix conditions. Verbs were monosyllabic when combined with the suffix [-t] and disyllabic when combined with the suffix [-n]. As can be seen in Table 1, the pronoun-by-suffix combinations created four conditions. 20 native German speakers were paid 10€ for their participation in this experiment. Recordings were performed in the sound attenuated chamber at the University of Tübingen. Midsagittal tongue images during articulating the target words were recorded using ultrasound imaging (Articulate Assistant Advanced (AAA) (Articulate Instruments Ltd. 2012)). The transducer was placed under the chin and kept in place by means of a headset.

Table 1: Four combinations of pronouns and suffixes with an example verb *malen* ([maːl(ə)n] "paint"). Combinations between the vowel in the pronoun and the suffix are created as tags for the conditions (highlighted in bold).

	Suffix	
Pronoun	[-t]	[-n]
[-iː]	**[-iː-aː-t]** sie malt [ziː maːlt]	**[-iː-aː-n]** sie malen [ziː mal(ə)n]
[-iɐ]	**[-iɐ-aː-t]** ihr malt [iɐ maːlt]	**[-iɐ-aː-n]** wir malen [viɐ mal(ə)n]

2.2. Preprocessing data

To normalize for different sizes of oral cavity among participants, ultrasound frames were cropped at the following boundaries (cf. Figure 1): (A) the border between the skin part on the bottom and the mylo-/genio-hyoid muscles, (B) the right edge of the hyoid shadow in the left side of images, (C) the maximum height of the tongue body during [k], and (D) the left edge of the mandible shadow in the right side of images. These

cropping-points were determined for each participant.

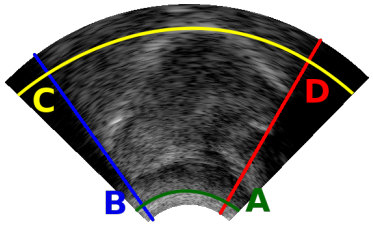


Figure 1: Cropping points with the mouth front to the right.

To reduce the size of the data and thus estimation time of fitting statistical models, the recorded ultrasound images were averaged pixel-wise across speakers for each pronoun-by-suffix condition. In each recording, five frames were selected within the stem vowel, so that they correspond roughly to T=0.00 (the onset of the stem vowel), 0.25, 0.50, 0.75, and 1.00 (the offset of the stem vowel). For each of these time points, the recorded ultrasound images were averaged pixel by pixel across participants in each pronoun-suffix combination.

2.3. Statistical analysis

Typically, ultrasound images are analyzed by spline curves fitted on tongue surface curves (Stone, Goldstein, and Zhang 1997; Davidson 2006; Dawson, Tiede, and Whalen 2016; Noiray et al. 2019) This method provides detailed information about tongue surface movements but misses considerable amount of information from other parts in ultrasound images.

In order to make use of as much information as possible in ultrasound images, we propose a different analysis method in the present study. We analyze the whole ultrasound image using Generalized Additive Mixed-effects Models (GAMMs, Wood 2011). To this end, the "pyult" python package was developed (Saito 2020). The package provides functions for preprocessing ultrasound images for fitting GAMMs in an R environment. Source code and details are available in https://github.com/msaito8623/pyult.

GAMMs can estimate non-linear relationships between a dependent variable and one or multiple numeric predictors. As the dependent variable in the present study, brightness values of pixels were adopted. Pixels' x- and y-coordinates were treated as predictors, in addition to an interaction between pronoun-by-suffix condition and frequency.

Individual models were fitted to the recorded ultrasound images at time steps T = 0.00 (the onset of the stem vowel), 0.25, 0.50, 0.75, 1.00 (the offset of the stem vowel).

2.4. Interpretation of GAMM Results

As explained above, we fitted the brightness values in the ultrasound images as a function of their x- and y-coordinates. Figure 2 demonstrates how ultrasound images are reflected in the GAMM analysis. Figure 2a and b illustrate the shape of the tongue surface in two pseudo-speakers. In Figure 2c, the averaged ultrasound image, which is averaged across two pseudo speakers, is presented. Note that the average picture is brighter where the trajectories from the two pseudo-speakers overlap, but dimmer where they diverge. The GAMM fit to the average image is illustrated in Figure 2d by means of a colored surface plot, where red color represents higher brightness, dark green to navy represent lower brightness, and yellow to light green represent middle brightness in the input ultrasound image.

(a) (b) (c) (d)

Figure 2: Example of how variance is reflected in GAMMs. (a & b) Pseudo-speaker 1 & 2. (c) Average between the two pseudo-speakers. (d) GAMM on (c). Red represents brighter and dark green to navy darker pixels in the input ultrasound image. Greater positional uncertainty results dimmer (green) colors in the GAMM.

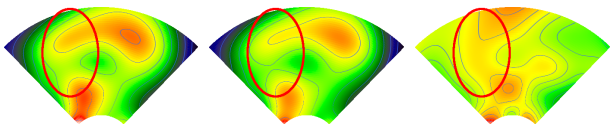
This example demonstrates that greater variance in tongue position between speakers is reflected by less bright pixels in the input ultrasound image (Figure 2c) and by wider and dimmer green areas in the surfaces plot (Figure 2d). This means that red in the surface plot represents low uncertainty about the tongue's position in the ultrasound image across speakers (and trials); wider and dimmer regions in the surface plot represent higher uncertainty about the tongue's position.

In the next section, we first discuss the effects of the pronouns and the suffixes on the tongue position, which we regard as controls of phonetic context on the articulation of [aː]. Ultrasound images for each phonetic context are obtained keeping the frequency value to its median. Then we present the effects of word frequency. For the presentation, we binned word frequency to high and low, which correspond to 90% and 10% quantile points.

3. Results

3.1. Coarticulation with pronouns and suffixes

The fan-shaped surface plots in Figures 3a and 3b illustrate the effects of carry-over coarticulation at the onset of the stem vowel (T = 0) from the ending vowels of the pronouns, i.e. [-iː] vs [-iɐ]. Figure 3c illustrates the difference.



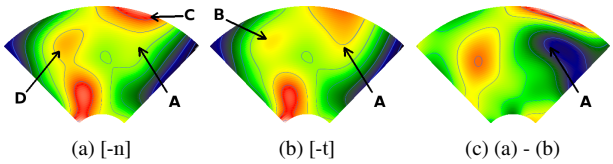
(a) [-iː] (b) [-iɐ] (c) (a) - (b)

Figure 3: Effects of carry-over coarticulation from the vowels (a) [-iː] and (b) [-iɐ] in the pronouns. (c) Difference between the two conditions.

We observe warmer colors in the [-iː] condition and less warmer colors in the [-iɐ] condition in the area of the tongue body/root (highlighted by red circle). This is supported by the difference plot (Figure 3c) that shows warmer colors in this region. Since the difference in brightness is in the same location in both pronoun conditions, this indicates that the uncertainty among speakers about the location of the tongue is higher in the [-iɐ] condition than in the [-iː] condition (cf. Figure 2). In other words, when the verb is preceded by 'ihr' [iɐ], there is more variability in the tongue body/root at the onset of [a(ː)] than when the verb is preceded by 'sie' [ziː].

Figure 4 demonstrates the effects of anticipatory coarticu-

lation of the suffixes [-t] and [-n] at the offset of stem vowels (T=1.00).



(a) [-n] (b) [-t] (c) (a) - (b)

Figure 4: Effects of anticipatory coarticulation (a) [-n] and (b) [-t] suffixes. (c) Difference between the two conditions.

We observe the warmer-colored region extending downward in the [-t] condition ('A' in Figure 4b) than in the [-n] condition (Figure 4a), supported by the strong difference in the difference plot. By contrast, the tongue tip/blade region is focused towards the palate in the [-n] condition ('C' in Figure 4a).

The differences in the tongue tip/blade region are reflected by differences in the tongue body/root region. The tongue body/root region is less warmer in the [-t] condition ('B') than in the [-n] condition, indicating higher uncertainty about the tongue's location in the [-t] condition.

The differences between the conditions indicate that there is a higher certainty that speakers placed their tongue tip at the palate when anticipating upcoming [-n] than when anticipating upcoming [-t]. This finding furthermore indicates that there is a greater variability between speakers in the anticipation of [-t] than in the anticipation of [-n].

3.2. Frequency effect

Among the time steps each of which a GAMM model was fitted to, T=0.75 showed the most pronounced effect of the frequency effect. Due to the lack of space, we focus on this time step in the present paper.

Effects of word frequency are illustrated in Figure 5 with the different pronoun-by-suffix conditions as different rows. High and low frequencies are in the first and second columns. Color coding is the same as in Figure 2.2. The third column represents the differences between high and low frequency, obtained by subtracting the estimates for low frequency from the estimates for high frequency.

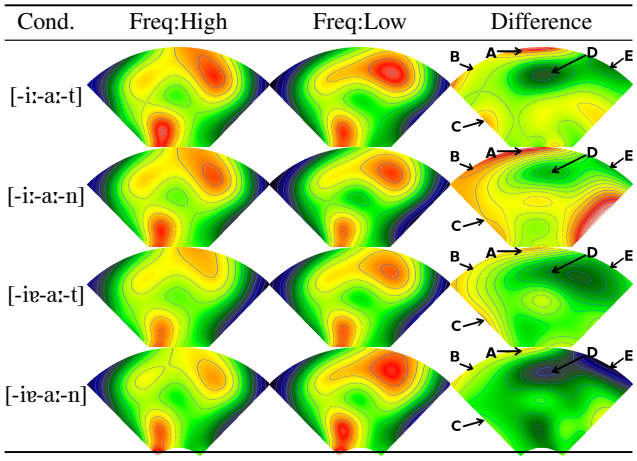


Figure 5: Fitted ultrasound images at T=0.75.

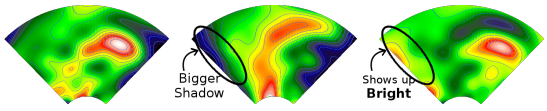
In the difference plot, 'A' represents roughly the location of the tongue middle, 'B' that of the tongue back, and 'C' that of the tongue root. As can be seen, the regions 'A', 'B', 'C' appear in warmer colors, indicating that the corresponding regions are in warmer colors in high frequency than low frequency.

The regions 'D' and 'E' appear in darker colors, indicating that the estimated images for high frequency words are darker in these regions. These changes in brightness indicate systematic differences in tongue position between high and low frequency words, independent of the pronoun-by-suffix conditions.

Warmer colors in the regions 'A' and 'B' in images for high frequency words indicate that the tongue middle to back tends to be higher in high frequency words than in low frequency words. The difference in height is also reflected by darker colors in the difference plots at the region 'D'. Furthermore, warmer colors in the difference plots in the region 'C' indicate that the tongue root was retracted in high frequency words and fronted in low frequency words.

Finally, we observe darker colors in the region 'E'. Darker colors in the difference plots represent brighter pixels in the corresponding region in the estimated ultrasound images of low than high frequency. Since the region 'E' is located roughly at the tongue tip, the region 'E' is indicating the tongue tip tends to be higher and closer to the palate in low frequency than in high frequency.

Figure 6 highlights these differences between high and low frequencies. For this, the first author recorded two tongue positions, one in resting position (6a) and the other when the tongue tip was pushed forward (6b). As can be seen in Figure 6b, the fronting movement of the tongue pulls the hyoid shadow (the shadow in the left side of ultrasound images) also forward (towards the right side of the images). The appearance of the shadow is reflected by darker colors in the area of the tongue back in Figure 6b, which is in turn mirrored by the corresponding region with warmer colors in the difference plot (Figure 6c). At the same time, the tongue tip is lifted slightly to be pushed forward, which creates the dark region in the top right corner of the difference plot (Figure 6c).



(a) (b) (c)

Figure 6: (a) Tongue in the resting position, (b) when fronted, and (c) the difference between (a) and (b).

4. Discussion

The present study aimed at replicating and extending the findings by Tomaschek, Tucker, et al. (2018), who investigated the effect of word frequency on articulatory trajectories in [aː] stem vowels in German verbs by means of electromagnetic articulography. They reported greater clarity for low frequency words and smoother tongue trajectories for high frequency words during anticipatory coarticulation of [aː] with suffixes. The present study used ultrasound to investigate the effect of word frequency while controlling for carry-over and anticipatory coarticulations.

The present study observed greater uncertainty about the location of tongue back/root at the onset of the stem vowel for the pronoun condition [-iɐ] than [-iː]. Furthermore, while anticipatory coarticulation is observed in both suffix conditions ([-n] and [-t]), the tongue tip/blade positions were more vari-

able when the upcoming suffix was [-t]. In line with the DIVA model (Guenther 2016), these results indicate different degrees of variability for an articulatory gesture ([aː]) depending on the context. According to the DIVA model, articulatory movements aim for targets defined by a higher-dimensional auditory and somatosensory maps. As a result of the experience with varying context, these maps differ in the size of the potential target regions. For example, while [t] allows different landing places along the palate, it is highly restricted in its vertical position. DIVA acknowledges that due to experience, sequences of multiple gestures can form a chunk, thus accounting for learning.

Accordingly, and in line with Tomaschek, Tucker, et al. (2018), the present study observed systematically different coarticulation strategies at the offset of the stem vowel in words of high and low frequency. The study finds higher tongue tip, lower tongue middle and back, in addition to fronted tongue root in low frequency words as compared to high frequency words (Figure 5). The shape of the tongue for low frequency words is closer to the canonical articulation of alveolar phones, in which the tongue tip is actively involved and raised to touch the ceiling, attempting clearer articulation (Figure 7).

Typically, effects of frequency (and probability) are interpreted as effects of informativity associated with reduction (Aylett and Turk 2004; Jaeger 2010). Due to the complex pattern of articulatory strategies, the present results do not support only reduction in relation to higher frequency. Rather, we interpreted these results in line with Tomaschek, Tucker, et al. (2018) and Tomaschek, Arnold, et al. (2018). They indicate that word frequency represents a measure of accumulated practice on a particular sequence of articulatory gestures, which is reflected by smoother tongue trajectories while enabling faster and more complex movements when necessary. In other words, practice makes perfect.

Figure 7: Schematized comparison between tongue shapes in low and high frequencies.

5. Acknowledgments

This study is funded by the Deutsche Forschungsgemeinschaft (Research Unit FOR2373 'Spoken Morphology', Project 'The articulation of morphologically complex words (BA3080/3-2)').

6. References

Articulate Instruments Ltd. (2012). *Articulate Assistant Advanced User Guide: Version 2.14.* Edinburgh, UK.

Aylett, M. and A. Turk (2004). "The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". In: *Language and Speech* 47.1, pp. 31–56.

Davidson, Lisa (2006). "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance". In: *The Journal of the Acoustical Society of America* 120.1, pp. 407–415.

Dawson, Katherine M., Mark K. Tiede, and D. H. Whalen (2016). "Methods for quantifying tongue shape and complexity using ultrasound imaging". In: *Clinical Linguistics and Phonetics* 30.3-5, pp. 328–344. DOI: 10.3109/02699206.2015.1099164.

Dell, G. S. (1986). "A spreading-activation theory of retrieval in sentence production". In: *Psychological review* 93.3, pp. 283–321.

Gahl, Susanne (2008). "Time and Thyme Are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech". In: *Language* 84.3, pp. 474–496.

Guenther, Frank H. (July 2016). *Neural Control of Speech*. en. MIT Press.

Jaeger, T. F. (2010). "Redundancy and reduction: Speakers manage syntactic information density". In: *Cognitive Psychology* 61.1, pp. 23–62.

Levelt, Willem J. M., Ardi Roelofs, and Antje S. Meyer (1999). "A theory of lexical access in speech production". In: *Behavioral and Brain Sciences* 22, pp. 1–75.

Levelt, Willem J. M. and Linda Wheeldon (1994). "Do speakers have access to a mental syllabary?" In: *Cognition* 50, pp. 239–269.

Lindblom, B. (1990). "Explaining Phonetic Variation: A Sketch of the H&H Theory". English. In: *Speech Production and Speech Modelling*. Ed. by Alain Marchal and WilliamJ. Hardcastle. Vol. 55. Springer Netherlands, pp. 403–439.

Noiray, Aude, Martijn Wieling, Dzhuma Abakarova, Elina Rubertus, and Mark Tiede (2019). "Back from the future: non-linear anticipation in adults and children's speech". In: *Journal of Speech, Language and Hearing Research*.

Öhman, S. E. G. (1966). "Coarticulation in VCV Utterances: Spectrographic Measurements". In: *The Journal of the Acoustical Society of America* 39, pp. 151–168. DOI: 10.1121/1.1909864.

Pluymaekers, Mark, Mirjam Ernestus, and R Harald Baayen (2005). "Lexical frequency and acoustic reduction in spoken Dutch". In: *The Journal of the Acoustical Society of America* 118.4, pp. 2561–2569. DOI: 10.1121/1.2011150.

Roelofs, Ardi (1997). "The WEAVER model of word-form encoding in speech production". In: *Cognition* 64.3, pp. 249–284.

Saito, Motoki (2020). *Pyult: Preprocessing utilities for ultrasound data in Python*. DOI: https://doi.org/10.5281/zenodo.4022838.

Song, Jae Yung, Katherine Demuth, Stefanie Shattuck-Hufnagel, and Lucie Ménard (2013). "The effects of coarticulation and morphological complexity on the production of English coda clusters: Acoustic and articulatory evidence from 2-year-olds and adults using ultrasound". In: *Journal of Phonetics* 41.3-4, pp. 281–295.

Sosnik, R., B. Hauptmann, A. Karni, and T. Flash (2004). "When practice leads to co-articulation: the evolution of geometrically defined movement primitives". In: *Exp Brain Res* 156, pp. 422–438.

Stone, Maureen, Moise H. Goldstein, and Yongqing Zhang (1997). "Principal component analysis of cross sections of tongue shapes in vowel production". In: *Speech Communication* 22.2-3, pp. 173–184.

Tomaschek, Fabian, Denis Arnold, Franziska Bröker, and R. Harald Baayen (2018). "Lexical frequency co-determines the speed-curvature relation in articulation". In: *Journal of Phonetics* 68, pp. 103–116.

Tomaschek, Fabian, Benjamin V. Tucker, Matteo Fasiolo, and R. Harald Baayen (2018). "Practice makes perfect: the consequences of lexical proficiency for articulation". In: *Linguistics Vanguard* 4.

Tomaschek, Fabian, Martijn Wieling, Denis Arnold, and Harald Baayen (2013). "Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience". In: *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, pp. 1302–1306.

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society (B)* 73, pp. 3–36.