Vowel articulation affected by word frequency

Fabian Tomaschek[1], Benjamin V. Tucker[2],
Martijn Wieling[3], R. Harald Baayen[1]

[1]*Department of Linguistics, Quantitative Linguistics, University of Tuebingen, Germany*
[3]*Department of Linguistics, University of Alberta, Canada*
[3]*Department of Information Science, University of Groningen, The Netherlands*
fabian.tomaschek@uni-tuebingen.de, btucker@ualberta.com,
wieling@gmail.com, harald.baayen@uni-tuebingen.com

Abstract

A frequently replicated finding is that the frequency of words affects their phonetic shape. In English, high frequency words have been shown to contain more centralized vowels than low frequency words. By contrast, a recent study on vowel articulation in German has shown a contrary finding. At the gestural level, tongue movements in HF words showed more extensive vowel targets and less coarticulationwith consonants. This paper further evaluates the later finding by taking into account a large set of verbs covering the continuum between high and low frequency. In addition to frequency the effects of two factors were analyzed: inflection (sagt vs. sagen) and speech rate (normal vs. fast). Our results imply that language experience increases the proficiency with which words are articulated: speakers are able to plan and target tongue movements earlier.

Keywords: articulography, vowels, word frequency, learning

1. Introduction

It is well known that how often a word is pronounced affects its phonetic form. In contrast to low frequency words (LF), high frequency (HF) words tend to have a lower number of segments (Zipf 1935) and shorter acoustic durations (Gahl 2008) as well as a higher probability of deleting a segment (Aylett and Turk 2004; Munson and Solomon 2004). At the segmental level, acoustic data indicates that vowels in English HF words form a more contracted and centralized vowelspace than vowels in LF words (Munson and Solomon 2004; Munson 2007). Zipf (1935) explained such reduction processes by he principle of least effort. The more often an articulatory movement is executed, the more efficient it is performed insofar as 'unnecessary' movements are ommitted. Aylett and Turk (2004) in light of their Smooth Signal Redundancy Hypothesis -*SSRH*- further show that higher contextual predictability of a word increases the reduction process.

In Tomaschek et al. (2013), we investigated frequency effects at the articulatory level in German. We found that HF produced stonger peripheral articulation in [i] vowels and less coarticluation in [a] vowels. We reasoned that these effects indicate another important process in language experience: learning. Higher frequency not only enables the speaker to articulate more efficiently but also more precisely. We can consider this effect as another mode of articulation, which we want to call the *articulatory proficiency theory -APT-*. The present study is a follow up of Tomaschek et al. (2013) further investigating APT in words containing [aː]. Unlike Tomaschek et al. (2013), who used words from the extremes of the continuum in a categorical way, we increased the number of words along the entire frequency continuum.

In addition to frequency, we introduced the following two factors: speech rate and morphology by means of inflection. It has been shown that faster speech rate leads to a stronger temporal contraction (Hoole, Mooshammer, and Tillmann 1994) and centralization (Moon and Lindblom 1994) of vowels. The SSRH predicts that vowels in HF words produced at a fast speech rate should show the most reduction. By contrast, the APT predicts that vowels in HF words will show less reduction due to a fast speech rate than those in LF words as learning should enable the speaker to counteract reduction due to fast speech.

The manipulation of the inflection addresses the question of how inflected forms, especially regulars, are stored in the mental lexicon. Are they generated by rules that transform a stem from the lexicon? Or are they stored independently as suggested by studies with lexical decision tasks (Milin, Filipovic Durdevic, and Moscoso del Prado Martin 2009). For example, Stemberger and MacWhinney (1986) investigated the occurrence of speech errors by having subjects pronounce past tense forms of regular verbs. They found that regular HF verbs were less prone to errors than LF verbs. On the one hand, their results indicate that inflected forms, even regulars, are stored in the lexicon. On the other hand, they show that the experience with specific words results in improved mastery of those words. We hypothesize that if the inflected form is stored in the lexicon, higher frequency should facilitate the planning/articulation of these forms, insofar that the articulation of the inflection is anticipated and produced earlier.

2. Stimuli and Methods

2.1. Stimuli

27 German verbs were used as stimuli. Their first syllable was stressed and contained the phonologically long [aː] vowel (see Table 1). In the disyllabic condition, words were produced in a "sie ...." *they* .... context: *sie zahlen* [ziː tsaːlən], *sie mahnen* [ziː maːnən]. Nine of these were monosyllabic inflected forms produced in a "ihr ...." *you pl.* .... context: *ihr zahlt* [iːɐ tsːalt], *ihr mahnt* [iːɐ maːnt]. We used the logarithmic relative frequency (henceforth *frequency* or *log(P)*) of a word in the SDEWAC corpus (Shaoul and Tomaschek 2013). In addition, the consonants before and after the vowel were controlled for place of articulation: coronal-V-coronal, coronal-V-labial or labial-V-coronal.

Table 1: *Stimulus material, ordered by frequency:* C-C = place of articulation of consonants next to the vowel: coronal (c) or labial (l). log(P) = logarithm of relative frequency. Monosyllabic words are written in italics.

Word	C-C	log(P)	Word	CVC	log(P)
zahlen	c-c	-14.20	waten	l-c	-21.35
schlafen	c-l	-15.53	*lahmt*	c-l	-21.49
zahlt	c-c	-15.79	labern	c-l	-21.51
schaden	c-c	-16.59	faseln	l-c	-22.06
baden	l-c	-17.99	schaben	c-l	-22.24
mahnt	l-c	-18.43	latschen	c-c	-22.36
blasen	c-c	-19.07	*schlaft*	c-l	-22.45
bahnt	l-c	-19.14	*schabt*	c-l	-23.18
bahnen	l-c	-19.25	tafeln	c-l	-23.35
mahnen	l-c	-19.31	*latscht*	c-c	-23.72
stapeln	c-l	-19.66	*blast*	c-c	-24.91
fahnden	l-c	-20.42	*zahnt*	c-c	-25.03
tadeln	c-c	-20.60	zahnen	c-c	-25.24
lahmen	c-l	-20.80	–	–	–

2.2. Recording method

All recordings were conducted in a sound proof booth at the Department of Linguistics of the University of Tübingen. A total of 17 native German subjects (mean age: 26, sd = 3) were instructed to read the stimuli aloud after being presented on a computer screen. Each word in each context was presented once. The list was pseudo-randomized for each participant and divided into three parts. Each part was presented once in a *slow* (inter-stimulus-time: 600 ms; presentation-time: 800 ms) and once in a *fast* speaking condition (inter-stimulus-time: 300 ms; presentation-time: 450 ms).
Articulatory movements of the tongue were recorded with the NDI wave articulograph at a sampling frequency of 100 Hz. Simultaneously, the audio signal was recorded (Sampling rate: 22.05 kHz, 16bit) and synchronized with the articulatory recordings. To correct for head movements and to define a local coordinate system, a reference sensor was attached to the subjects' forehead. Before the tongue sensors were attached, a recording was made to determine the rotation from the local reference to a standardized coordinate system. The standardized coordinate system was defined by a bite plate to which three sensors in a triangular configuration were attached. Tongue movements were captured by three sensors: one slightly behind the tongue tip (TT), one at the tongue middle (TM) and one at the tongue body (TB; distance between each sensor: around 2cm). The present analysis focuses on the sensors TT and TB.

2.3. Preprocessing

The recorded positions of the tongue sensors were centered at the midpoint of the bite plate and rotated in such a way that the front-back direction of the tongue was aligned to the x-axis with more positive values towards the front of the mouth, and more positive z-values towards the top of the oral cavity. No filtering was applied as this would artificially increase the autocorrelation of the data. To determine segment boundaries, the audio signal was automatically aligned with phonetic transcriptions by means of a Hidden-Markov-Model-based forced aligner for German (Rapp 1995). Alignments were manually verified and corrected where necessary. The beginning (henceforth *CV transition*) and offset time points (henceforth *VC transition*) of each vowel in every word were used to identify the movement trajectories of the three tongue sensors.

3. Analysis and results

3.1. Analysis

Since the duration of each vowel differs from utterance to utterance per person and word, vowel duration was normalized between 0 and 1 (henceforth called *time*). Separate analysis of the vertical movement as a function of time in each of the sensors (TT, TB) was performed by means of generalized additive models (GAMs) (R version 3.0.2, package *mgcv*, Version 1.7-28, Wood 2006). GAMs model the nonlinear relationships between the numeric predictor and the response variable with thin plate regression splines.
Interactions between two gradual predictors are modelled by means of tensor product smooths with cubic spline basis functions and result in *wiggly* surfaces. The estimated degrees of freedom (*edf*) reflect the number of parameters required for modeling a wiggly curve, surface or hypersurface and measure how wiggly it is. More wiggly curves, surfaces or hypersurfaces require more edfs.
As the exact tongue movements might differ across subjects due to different morphologies of the oral cavity, by-subject random factor smooths as a function of time were included. In order to account for random item effects, random factor smooths for time per word were included. These random factor smooths have the same function as the combination of random intercepts and random slopes in a standard linear mixed-effects regression analysis. Random factor smooths were also included for CV and VC consonant place of articulation. Random wiggly curves were significant for the variation by participants, words and place of articulation in CV and VC consonants (Tables 2 and 3).
Articulatory data constitute time series with strong autocorrelation (i.e. one can predict the value in X+1 given the value in X). Residual autocorrelation results in anti-conservative p-values. We therefore included a parameter rho to remove autocorrelation noise from residuals. Remaining errors were Gaussian and uncorrelated. Autocorrelation was estimated on the basis of the first model. This first estimate was used during model optimization. After the optimal model was found, autocorrelation was estimated anew and optimized.
Model selection was based on model comparison with the maximum likelihood (*ML*) test. The optimal models are presented in Table 2 and 3. A detailed description of this approach can be found in Kryuchkova et al. (2012).

3.2. Results

3.2.1. Contour plots

We use contour plots to show fitted regression surfaces. For example, the contour plot in Fig. 1 displays the tongue tip sensor's vertical position (in mm) as a function of time (x-axis) and as a function of frequency (y-axis). Lighter shades indicate high positions, darker shades indicate low positions. Contour lines connect points with equal elevation. The movement for a certain probability is represented in the vertical axis. For example, the tongue tip's movement at log(P) = -20 starts at a height of 4 mm above the mean, falls to -2 mm (i.e. 2 mm below mean) and then rises to 4 mm again.

3.2.2. The tongue tip sensor (TT)

The model for the movement in the TT sensor yields an R-squared of 0.791 and explains 79.5% of the deviance in the data (ML: 24526, edfs: 388). No significant effects of speech rate and morphological alternation were found. The tensor model indicates that the time-by-frequency interaction is significant (Table 2, first row). In Fig. 1, one can see that the time point of the vowel target changes minimally as a function of frequency. Furthermore, the CV transition onset is higher at log(P) = -24 than at higher frequencies. Also, the VC transition raises earlier at this frequency than at higher frequencies. In all, the effect of frequency was tiny. Neither the falling and rising pattern of the tongue tip, nor the depth of the vowel target was drastically affected by frequency.

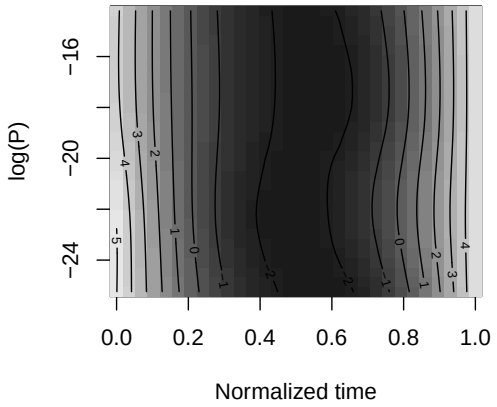

Figure 1: *Vertical position of the tongue tip sensor (TT) as a function of time (x-axis) and logarithmic relative frequency (y-axis):* Lighter shades indicate high positions, darker shades indicate low positions. Contour lines connect points with equal elevation.

Table 2: *GAM table for TT*

Effect	edf	F	p-value
tensor(time, frequency)	12.16	80.54	<0.0001
smooth(time, participant)	141.50	54.47	<0.0001
smooth(time, onset place)	8.67	26.43	<0.0001
smooth(time, offset place)	8.80	53.65	<0.0001
smooth(time, word)	215.83	10.61	<0.0001

3.2.3. The tongue body sensor (TB)

The model for TB sensor yields an R-squared of 0.40 and explains 41.6% of the deviance in the data (REML-score: 20358, edfs: 314). No significant effect of speech rate was present. The model indicates a significant morphology-by-frequency interaction (Table 2 first and second row; Fig. 2). In the disyllabic condition (stem + /en/), the vowel is produced lower with increasing frequency. Simultaneously, the CV and VC transitions become steeper. In the monosyllabic condition (stem + /t/), the movement pattern is reversed. With higher frequency, the vowel target becomes higher and the onset of the VC transition starts earlier.

Table 3: *GAM table for TB*

Effect	edf	F	p-value
tensor(time, frequency: '2syll.')	8.51	22.87	<0.0001
tensor(time, frequency: '1syll.')	15.64	26.12	<0.0001
smooth(time, participant)	136.35	23.91	<0.0001
smooth(time, onset place)	7.04	6.71	<0.0001
smooth(time, offset place)	6.86	5.22	<0.0001
smooth(time, word)	137.30	1.49	<0.0001

4. Discussion and conclusion

The present study investigated the effects of morphological inflection, speech rate and frequency of occurrence on the articulation of the German vowel [aː]. Two sensor positions, tongue tip and tongue body, were analyzed. The present findings indicate that none of the factors under investigation affected the movement pattern of the tongue tip. One possible reason for this finding might be that [aː] is primarily articulated by the body of the tongue whereas the tongue tip is used primarily for the production of coronal consonants. It is possible that word frequency effects on the CV and VC transitions were confounded by the articulation of the surrounding consonants. As Munson (2011) has shown that more frequent two-consonant clusters are articulated shorter than less frequent ones, phonotactic frequencies would be probably a better measure to investigate usage effects in the tongue tip.

Hoole, Mooshammer, and Tillmann (1994) report that faster speech rate results in a contraction of the vowel. However, we found no effect of speech rate at any of the sensors under investigation. This might be a result of the present analysis technique which required normalization of vowel duration between 0 and 1. Possible contractions might have been normalized out.

Frequency affected the movement patterns of the tongue body. In the monosyllabic condition, the [aː] vowel was produced with a less extensive, i.e. a higher and more centralized target. Simultaneously, the outgoing VC transition became smoother. This finding seems in line with the SSRH (Aylett and Turk 2004) and the principle of least effort (Zipf 1935), which state that the more frequent a certain word is, the less effort is invested into producing it. This is why the vowel would be centralized and stronger coarticulated with the consonantal context (Munson and Solomon 2004; Munson 2007).

In the disyllabic condition, the opposite pattern is visible. With increasing frequency, the vowel target is articulated more extensively and the CV and VC transitions become steeper. Steeper transitions indicate less coarticulation between vowel and consonant, as has been show by Katz and Bharadway (2001). This finding replicates Tomaschek et al. (2013).

How is it possible that the same vowel is affected in two different ways by frequency, depending on whether it is produced in a monosyllabic or in a disyllabic word? One could argue that in the disyllabic condition, there is no need to reduce the vowel in the first syllable. Rather, reduction occurs in the second unstressed syllable where the [ə] is often non-existent in modern German and the sonorant becomes syllabic (Becker 1998). Since in the monosyllabic condition there is no unstressed syllable where reduction might be focussed, it is realized on the vowel itself.

However, we would like propose another explanation for this

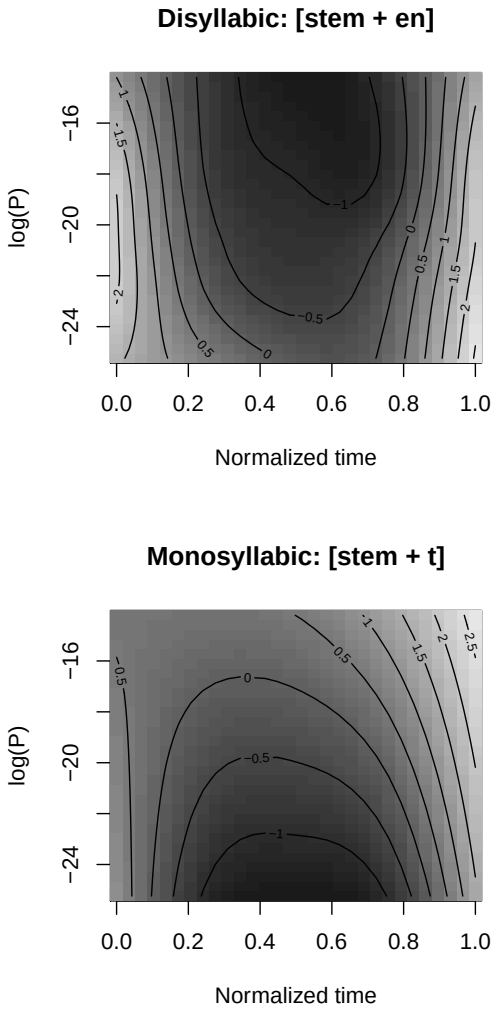

Figure 2: *Vertical position of the tongue body sensor as a function of time (x-axis) and frequency (y-axis):* First row: monosyllabic condition. Second row: disyllabic condition.

finding. In the monosyllabic condition, a [t] had to be attached to the stem. In order to produce the [t], the tongue body raises so that the occlusion which occurs between the VC consonant and the [t] can be produced. We observed that the tongue body raises earlier, the higher the frequency. This early raising implies earlier preparation of the articulation of the [t]. Given word frequency as a measure of experience, our results indicate a learning effect in both conditions: In the disyllabic condition, the vowel shows less coarticulation with increasing frequency; in the monosyllabic condition, the final segment is anticipated and targeted earlier.

In summary, our data show that word frequency does not affect the movement patterns of the tongue tip. Frequency manifest itself in the tongue body, where it interacts with the word's inflectional form. Our articulatory proficiency theory *APT* captures this phenomenon insofar as acoustically measured reductions at the spectral and temporal level turn out to be an earlier preparation of articulatory movements. Nevertheless, on the basis of the present data, no conclusion can be drawn and open questions have to be solved in future studies.

5. Acknowledgements

We would like to thank our student assistants Franziska Bröker, Dankmar Enke, Lea Hofmaier and Samuel Thiele for their inexhaustible willpower while correcting the annotations. This paper was funded by the Alexander von Humboldt Professorship.

6. References

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.

Becker, T. (1998). *Das Vokalsystem der deutschen Standardsprache ("The Vowel System of the German Standard Language")*. Frankfurt am Main: Peter Lang.

Gahl, S. (2008). ""Thyme" and "Time" are not homophones. Word durations in spontaneous speech". In: *Language* 84.3, pp. 474–496.

Hoole, P., C. Mooshammer, and H.G. Tillmann (1994). "Kinematic analysis of vowel production in German". In: *Proceedings of ICSLP 94, Yokohama*, pp. 53–56.

Katz, W.F and S. Bharadway (2001). "Coarticulation in fricative-vowel syllables produced by children and adults: a preliminary report". In: *Clinical linguistics and phonetics* 15.1, pp. 139–143.

Kryuchkova, T., B.V Tucker, L.H. Wurm, and R.H. Harald (2012). "Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography". In: *Brain and Language* 122, pp. 81–91.

Milin, P., D. Filipovic Durdevic, and F. Moscoso del Prado Martin (2009). "The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian". In: *Journal of Memory and Language*, 50–64.

Moon, S-J. and B. Lindblom (1994). "Interaction between duration, context, and speaking style in English stressed vowels." In: *he Journal of the Acoustical Society of America* 96, pp. 40–55.

Munson, B. (2007). "Lexical access, lexical representation, and vowel production". In: *Laboratory Phonology* 9, 201–228.

— (2011). "Phonological pattern frequency and speech production in adults and children". In: *Journal of speech and hearing research* 44, 778–792.

Munson, B. and N.P. Solomon (2004). "The effect of phonological neighborhood density on vowel articulation". In: *Journal of speech and hearing research* 47, pp. 1048–1058.

Rapp, S. (1995). "Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German". In: *Proceedings of ELSNET goes east and IMACS Workshop*. Moscow.

Shaoul, C. and F. Tomaschek (2013). "A phonological database based on CELEX and N-gram frequencies from the SDEWAC corpus". In: *Personal communication*.

Stemberger, J.P. and B. MacWhinney (1986). "Frequency and the lexical storage of regularly inflected forms". In: *Memory and Cognition* 14.1, pp. 17–26.

Tomaschek, F., M. Wieling, D. Arnold, and H. Baayen (2013). "Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience". In: *Proceedings of the Interspeech, Lyon*.

Wood, S. (2006). Boca Raton, Florida, U. S. A: Chapman and Hall/CRC.

Zipf, G.K. (1935). *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. Cambridge: MIT Press.