INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Dimensions of prosodic prominence in an attractor model

Simon Roessig, Doris Mücke, Lena Pagel

I*f*L – Phonetik, Universität zu Köln
mail@simonroessig.de, {doris.muecke, lena.pagel}@uni-koeln.de

Abstract

Speakers of intonation languages use bundles of cues to express prosodic prominence. This work contributes further evidence for the multi-dimensionality of prosodic prominence in German reporting articulatory (3D EMA) and acoustic recordings from 27 speakers. In particular, we show that speakers use specific categorical and continuous modifications of the laryngeal system (tonal onglide) as well as continuous modifications of the supra-laryngeal system (lip aperture and tongue body position) to mark focus structure prosodically. These modifications are found between unaccented and accented but also within the group of accented words, revealing that speakers use prosodic modulations to directly encode prominence. On the basis of these findings we develop a dynamical model of prosodic patterns that is able to capture the manipulations as the modulation of an attractor landscape that is shaped by the different prosodic dimensions involved.

Index Terms: Intonation, articulation, prosodic strengthening, dynamical systems, attractors, prominence

1. Introduction

In intonation languages, a multitude of prosodic cues can be modulated to express prosodic prominence. Speakers use strategies involving laryngeal modifications like the placement of a pitch accent, the choice between different pitch accent types, and the manipulation of phonetic parameters of pitch accents like target height. Studies have shown that rising pitch accent types convey more prominence than falling accents [1], and that within the group of rising accents, larger rises serve to express higher prominence compared to shallower rises [2, 3, 4]. In addition, speakers use strategies that involve the supra-laryngeal system to convey prosodic prominence. One strategy has been discussed under the term of *sonority expansion*. To make a syllable more prominent, speakers produce louder and more sonorous syllables by opening the mouth wider as an open oral cavity allows for a greater radiation of acoustic energy from the mouth. Another strategy is referred to as *localised hyperarticulation* [5]. It is based on the H&H model developed by [6] and follows the observation that prominent parts of an utterance exhibit a more extreme articulation of the tongue body in vowel productions. This strategy serves to make speech sounds more distinct: A low vowel such as /a/, for example, shows a lower position of the tongue body, while a front vowel such as /i/ shows a more fronted tongue body position [7, 8]. Many of the studies investigating prosodic strengthening of articulation in prominent positions have focussed on the dichotomy between unaccented and accented syllables. More recently, [9] presented data on focus marking in German showing that there are not only adjustments of lip opening gestures between accented and unaccented syllables but also within the group of accented syllables reflecting the focus structure of the phrase directly.

The first aim of the present study is to investigate prosodic modifications in both the tonal as well as the articulatory domain. In doing so, we will concentrate on modifications induced by accent *and* modifications found within the group of accented words. Our analysis includes acoustic and articulatory recordings of 27 German speakers producing words in background and in two different types of focus.

The results will be embedded in an attractor model building upon the theoretical framework of dynamical systems – a framework that has gained attention from many researchers in the last decades. Attractor models acknowledge the dynamical nature of the mind [10, 11] and are able to propose alternatives to overcome a strict separation of the categorical and the continuous aspects of cognition. More specifically, while phonology and phonetics have been conceptualised as separate domains – with phonology operating in a symbolic, discrete domain and phonetics operating in a continuous domain – a dynamical perspective of language and speech views them as a single system [12]. In this system, categorical stability is provided by relatively stable states that the system will gravitate to over time, called *attractors*. At the same time, being situated in a completely continuous space, the attractors allow for flexibility. This view resonates with the growing body of research that shows how important detailed phonetic variation is [13, 14]. In prosody research, dynamical systems offer a useful tool to account for the variability and flexibility of prosodic categories. In the present work, we sketch a dynamical system that accounts for the combination of modifications found in the expression of prosodic prominence. We will conceptualise prosodic prominence as a multi-dimensional attractor landscape that each individual prosodic dimension contributes to with different degrees of complexity.

2. Methods

Acoustic and articulatory recordings were carried out at the IfL Phonetics department at the University of Cologne using a head-mounted condenser microphone and a Carstens AG501 3D electromagnetic articulograph to track the movements of the articulators. We recorded 27 monolingual native speakers of German aged between 19 and 35 years (17 female). The subjects were prompted to produce the target utterances by involving them in an interactive animated game shown on a screen in front of them. In the game they were asked to help a robot to retrieve tools hidden in a factory. The robot's questions served as triggers for the focus structure of the answer, such that the target word (a fictitious object on which the tool is placed) was in broad, narrow or contrastive focus or in the background (only background, broad and contrastive are analysed here).

As target words 20 German sounding disyllabic nonce words with a C_1V_1:C_2ə structure were created. C_1 was chosen from the set of /n m b l v/, C_2 from /n m z l v/. The first, stressed vowel V_1 was either /a:/ or /o:/ and the second, unstressed vowel was always schwa. The order of trials was randomised for each participant. Example question-answer-pairs

http://dx.doi.org/10.21437/Interspeech.2019-2227

eliciting the three focus structures analysed in this paper are shown below ("Wohse" is the target word in these examples):

Background: Q: Hat er die Säge auf die Wohse gelegt?
Did he put the saw on the Wohse?
A: Er hat [den Hammer]$_F$ auf die Wohse gelegt.
He put the hammer on the Wohse.
Broad: Q: Was hat er gemacht?
What did he do?
A: Er hat [den Hammer auf die Wohse gelegt.]$_F$
He put the hammer on the Wohse.
Contrastive Q: Hat er den Hammer auf die Mahse gelegt?
Did he put the hammer on the Mahse?
A: Er hat den Hammer auf [die Wohse]$_F$ gelegt.
He put the hammer on the Wohse.

Using the emuR speech database system [15], a trained annotator labelled the start and end of the stressed syllable in the acoustic signal of each target word. Within the acoustic boundaries of this syllable, the kinematic data was extracted automatically. As an index for the lip aperture, we evaluated the maximum Euclidian distance between the lips [16]. Additionally, the lowest point of the tongue body was measured, i.e. the minimum of the recorded trajectory within the labelled syllable. All values for the lip distance and tongue body position were z-scored for every subject.

For the intonational analysis, we measured the tonal onglide of each nuclear accent which characterises the portion of the f0 movement towards the main tonal target of the pitch accent [17]. In terms of an autosegmental-metrical analysis [18], like GToBI [19], positive onglide values depict the rising f0 movement of L+H* and H* pitch accent types while negative onglide values represent the falling f0 movement of a L* or !H* accent. Besides capturing the direction of the tonal movement by distinguishing rising from falling, the tonal onglide also reflects the magnitude of the rise or the fall. Two trained labellers judged perceptually whether the nuclear accents on the target words were rising or falling. Using a consistent labelling scheme, they identified the beginning and the end of the f0 movement in a window of three syllables including the accented syllable as well as the preceding and following syllable. If the target word was unaccented, as it was the case in almost all utterances of the background condition, an alternative onglide measure was applied using fixed time points (5 ms before the start and 50 ms before the end of the stressed syllable). Onglide measures were normalised for individual speakers by dividing each value by the speaker mean for rising or falling accents respectively or by the overall absolute mean in case of an unaccented word. As stated above, this paper analyses a subset of the data including only the vowel /a/ and the focus conditions background, broad and contrastive.

3. Results

Figure 1 (top) presents the onglide distributions of all speakers for the three focus conditions as density plots. In the background condition, the data show a single mode located slightly below zero. When going from background to broad, i.e. from unaccented to accented, we can observe that the distribution changes to a bimodal shape. With the occurrence of accent, there are now two possibilities: falling and rising. In contrastive focus, the right mode is clearly more pronounced, only a small number of falling accents is produced in this focus condition. Since rising accents dominate the data, we look at the means of rising accents in figure 1 (bottom). Taken together, the following pattern can be attested: In addition to the increase in the number of accents with rising onglide, the rises become higher, as reflected in the growth of the mean from broad focus to contrastive focus.

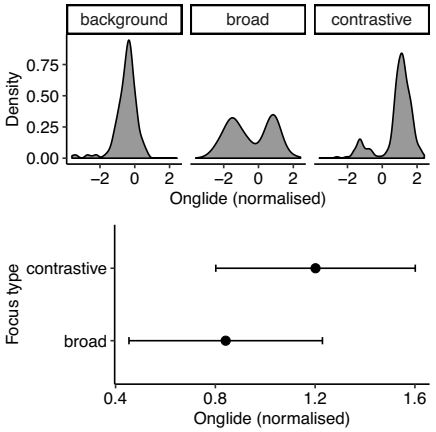

Figure 1: *Distribution of normalised onglide values (top) and means + standard deviations of rising onglide subdistributions (bottom).*

We analyse the results using a Bayesian linear mixed model in R [20] with brms [21] that implements an interface to Bayesian inference in Stan [22]. We report the estimated differences between focus conditions in terms of posterior means, 95% credible intervals, and the probability of the estimate being greater or smaller than 0 (depending on the relation we look at).

The model includes normalised onglide as the dependent variable, focus type as a fixed effect, and random intercepts for speakers and target words as well as by-speaker and by-target-word slopes for the effect of focus type. Since the distribution of the dependent variable is bimodal, we use a prior for the predictor that is characterised by a mixture of two gaussian distributions centred around -0.5 and 0.5 respectively. The model estimates the parameter theta that represents the extent to which the two gaussian distributions are mixed. For this parameter, we use a prior centred around 0. Differences in theta indicate the differences in the proportions of the two modes in the onglide data. The model runs with four sampling chains of 5000 iterations each, with a warm-up period of 3000 iterations.

Our investigation of the Bayesian model starts with the mixing parameter. To calculate the differences between focus types, we subtract the posterior samples for background from broad focus (broad–background) and broad focus from contrastive focus (contrastive–broad). Given the model and the data, the Bayesian analysis yields strong evidence for differences in the posterior probabilities for the mixing parameter between background and broad ($\hat{\beta} = -2.46, CI = [-4.15, -0.86], Pr(\hat{\beta} > 0) = 1$), as well as between broad and contrastive ($\hat{\beta} = 2.99, CI = [1.00, 5.03], Pr(\hat{\beta} < 0) = 1$). Note that the difference between broad and background is estimated to be negative but positive with contrastive-broad, because the model uses the right mode of the prior to model the single mode of background. This mode shrinks when going from background to broad and it grows again when going from broad to contrastive.

To assess the differences between the focus conditions regarding the rising distributions, we investigate the mean esti-

mates of the right gaussian sub-distribution. We only look at broad and contrastive focus since we can only talk of a rising accent in these conditions, background is unaccented. The model provides evidence for differences in the posterior probabilities between broad focus and contrastive focus ($\hat{\beta} = 0.37, CI = [0.19, 0.56], Pr(\hat{\beta} > 0) = 1$).

Figure 2 shows the means and standard deviations of the maximal Euclidean lip distance (top) and the minimal tongue body position (bottom). The results reveal the following pattern: From background to broad focus and from broad to contrastive focus, the inter-lip distance increases and the tongue body is lowered. As above, we analyse the data using Bayesian linear mixed models. Because the data is not bimodally distributed, we do not use a mixture prior here and thus do not estimate the mixing parameter. Maximal Euclidean lip distance or minimal tongue body position are included as dependent variables. The model consists of focus type as fixed effect and random intercepts for speakers and target words as well as by-speaker and by-target-word slopes for the effect of focus type.

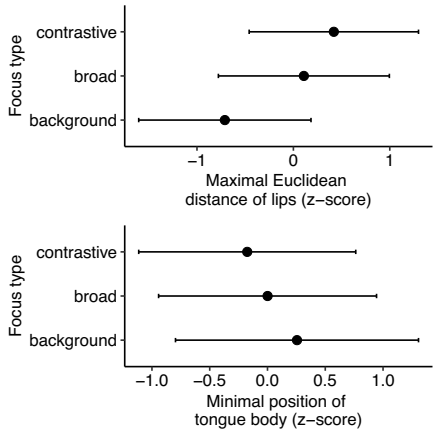

Figure 2: *Means and standard deviations for maximal Euclidean lip distance (top) and minimal position of the tongue body (bottom).*

Given the model and the data, the analysis yields clear differences in the posterior probabilities between background and broad focus ($\hat{\beta} = 0.81, CI = [0.65, 0.97], Pr(\hat{\beta} > 0) = 1$), and broad focus and contrastive focus ($\hat{\beta} = 0.30, CI = [0.13, 0.48], Pr(\hat{\beta} > 0) = 1$) regarding the lip distance measure. Likewise, the analysis provides evidence for differences in the posterior probabilities between background and broad focus ($\hat{\beta} = -0.25, CI = [-0.44, -0.06], Pr(\hat{\beta} < 0) = 1$) and broad focus and contrastive focus ($\hat{\beta} = -0.17, CI = [-0.38, 0.05], Pr(\hat{\beta} < 0) = 0.94$).

4. Modelling

The results presented in the previous section show the following pattern: On the tonal tier, when going from background to broad focus, i.e. unaccented to accented, the distribution of flat f0 is split into a bimodal distribution, representing the possibility of falling and rising accents. When going from broad focus to contrastive focus, the number of rising accents increases while the number of falling accents decreases. In addition, the rises become higher as a change from broad to contrastive focus. Both the dominance of rising accents as well as the increase in the magnitude of the tonal onglide of these rises make the focused word more prominent.

On the articulatory tier, we observe a continuous increase in the lip aperture and a lowering of the tongue body from background to broad focus and from broad to contrastive focus. The increase in inter-lip distance can be related to sonority expansion, i.e. the speaker intends to produce a louder vowel in the accented syllable [23, 8], strengthening the syntagmatic contrast between prominent and non-prominent syllables in the phrase. The lowering of the tongue during the low vowel /a/ can be related to the strategy of localised hyperarticulation. Both types of modification can be seen as strategies to enhance the prominence of the target word from background to contrastive focus with an intermediate step for broad focus. In what follows, we propose a dynamical system that models the tonal and articulatory modifications as the result of the scaling of one control parameter.

The dynamical perspective of the mind conceptualises the mind as being constantly in flux: a system that travels through its space of possible states on a smooth trajectory [11]. The attractors are the relatively stable states that the system gravitates towards. The dynamical system of states and changes of states can be captured using the language of differential equations [24]. In this formal language, one useful way of formulating a dynamical system is by giving its potential energy function. The graph of the potential energy curve gives a good impression of the attractors present in the system as the attractors surface as local minima in the potential energy curve. Two examples are shown in figure 3. On the left, the two-attractor landscape of the system with the potential energy function $V(x) = \frac{x^4}{4} - \frac{x^2}{2} - kx$ is presented. On the right, the attractor landscape is simpler with a single attractor, described by the potential energy function $V(x) = \frac{(x-k)^2}{2}$. The solid lines represent the attractor landscapes when the control parameter $k = 0$. When the control parameter is scaled to $k = 0.5$, the landscapes change as can be seen in the dashed lines: In the one-attractor landscape, the attractor basin moves towards higher values. In the two-attractor landscape, the basin of the right attractor becomes deeper. This means that this state becomes more stable and thus more probable. In addition, the deppest point of the attractor basin changes its location to slightly higher values.

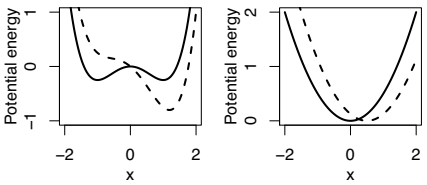

Figure 3: *Basic attractor landscapes with two (left) and one attractor (right). The dashed line represents the result of the scaling of the control parameter.*

We will use attractor landscapes similar to the ones presented to model the prosodic data. The double-well attractor landscape is suited to model the tonal onglide: In broad focus, the onglide data exhibits bistability whereas the right mode becomes more pronounced in contrastive focus – corresponding to a strengthening of the stability of the right attractor. This "tilt" to the right results in more rises and also higher rises. To model background, the model needs to be able to change to a single-attractor landscape as the control parameter is scaled,

since there is only one mode located roughly in the middle between the two modes for falling and rising accents. Such a dramatic change in a dynamical system is called a bifurcation [25, 26].

One possible system that unites all these features is given in equation 1. To model our data, we chose the following values for k for the focus conditions: background $k = -1$, broad $k = 1$, contrastive $k = 1.7$. The corresponding attractor landscapes are shown in figure 4. For background, only one attractor basin with the deepest point slightly below zero is present. As k is scaled up, the attractor landscape becomes bistable in broad and then tilts to the right side in contrastive. From an intonational perspective, as soon as the control parameter passes the critical value for accentuation, the system exhibits two possibilities. However, rising accents and larger positive tonal onglides become more frequent with higher k values.

$$V(x) = \frac{x^4}{4} - (1 - e^{-k})\frac{x^2}{2} - |k|(k-1)\frac{x}{4} \tag{1}$$

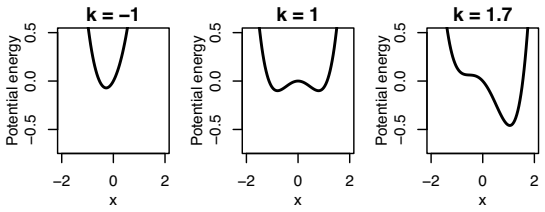

Figure 4: *Onglide attractor landscapes for background (left), broad (middle), and contrastive focus (right).*

In the results section, we have shown that in addition to the modulation of f0, the behaviour of the lingual and labial system is modified. We can view all these modifications as the outcome of a multi-dimensional system of prosody. In this system, the scaling of the control parameter results in a bundle of modifications. The attractors of the joint landscape come into being as a combination of these multiple dimensions. The way in which the dimensions shape the combined attractor landscape will be different: Some of the dimensions will contribute a rather complex shape, like the tonal onglide with its two stable states for falling and rising. Other dimensions will contribute a simpler shape, like the lip and tongue movements. Figure 5 ties together the dimensions of tonal onglide and lip distance to give a visual impression of a system operating with more than one prosodic dimension. For the lip distance, we assume a monostable attractor landscape in which the attractor basin can move to more extreme values as the control parameter is scaled (see figure 3 (right)). The same control parameter values as above are used in figure 5, each landscape is shown from two different angles (left and right each) to highlight the changes in the two dimensions.

When $k = -1$ (background), the whole landscape has a single attractor. As k is scaled to $k = 1$ (broad focus), the dimension of tonal onglide bifurcates and becomes bimodal, resulting in a two-attractor landscape. These attractors move "forward" to more extreme values on the lip distance axis. With $k = 1.7$, the right attractor basin becomes deeper, i.e. more stable, and the attractors move to more extreme values on both the tonal onglide as well as on the lip distance dimension. Of course, the inclusion of prosodic dimension does not have to stop here, only the limits of visualisation are reached. The tongue body contributes a dimension similar to the lip distance. With the scaling of k, the attractors move along this axis towards more extreme negative values.

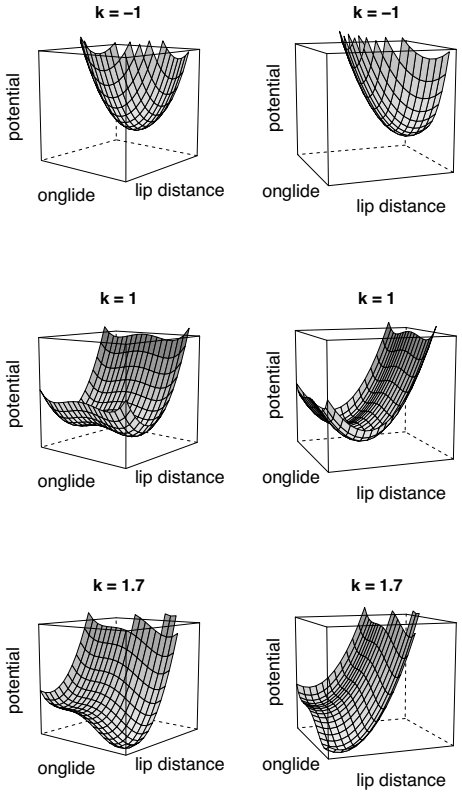

Figure 5: *Visualisation of two dimensions of the multi-dimensional prosody attractor landscape for background (top), broad (middle), and contrastive focus (bottom).*

5. Conclusion

We have shown that speakers of German modulate multiple parameters to express focus structure by means of prosodic prominence. On the tonal tier, the manipulations include categorical changes, like placement of a pitch accent or the choice between pitch accents, as well as continuous changes, like the adjustment of the onglide magnitude of the rises. On the articulatory tier, speakers open the lips wider to radiate more energy from the mouth. Furthermore, they lower the tongue body to produce a more distinct instance of the vowel /a/ in prominent positions. Crucially, these modifications do not only occur between unaccented and accented words but also within the group of accented words to increase prosodic prominence. We have sketched a model that describes these manipulations as changes in an attractor landscape that is shaped by the combined contribution of the individual prosodic dimensions.

6. Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of the SFB1252 "Prominence in Language" in the project A04 "Dynamic modelling of prosodic prominence" at the University of Cologne.

7. References

[1] S. Baumann and C. Röhr, "The perceptual prominence of pitch accent types in German," in *Proceedings 18th ICPhS, Glasgow, UK: University of Glasgow.*, aug 2015, p. 298.

[2] D. R. Ladd and R. Morton, "The perception of intonational emphasis: Continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[3] F. Kügler and A. Gollrad, "Production and perception of contrast: The case of the rise-fall contour in German," *Frontiers in Psychology*, vol. 6, p. 1254, 2015.

[4] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, "Integrating the discreteness and continuity of intonational categories," *Journal of Phonetics*, vol. 64, pp. 90–107, 2017.

[5] K. J. de Jong, "The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation." *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 491–504, jan 1995.

[6] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," in *Speech Production and Speech Modeling*, kluwer aca ed., W. Hardcastle and A. Marchal, Eds., Dortrecht, 1990, pp. 403–439.

[7] T. Cho and J. McQueen, "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," *Journal of Phonetics*, vol. 33, no. 2, pp. 121–157, 2005.

[8] J. Harrington, J. Fletcher, and M. E. Beckman, "Manner and place conflicts in the articulation of accent in Australian English," in *Papers in laboratory phonology V: Acquisition and the Lexicon*, M. Broe, Ed. Cambridge: Cambridge University Press, 2000, pp. 40–51.

[9] D. Mücke and M. Grice, "The effect of focus marking on supralaryngeal articulation - Is it mediated by accentuation?" *Journal of Phonetics*, vol. 44, pp. 47–61, 2014.

[10] R. Port, *Dynamical Systems Hypothesis in Cognitive Science*. Nature Publishing Group, 2002, pp. 1027–1032.

[11] M. Spivey, *The Continuity of Mind*. New York: Oxford University Press, 2007.

[12] A. I. Gafos, "Dynamics in Grammar," in *Laboratory phonology 8: Varieties of Phonological Competence*, M. L. Goldstein, D. H. Whalen, and C. Best, Eds. Berlin, New York: Mouton de Gruyter, 2006, pp. 51–79.

[13] J. Pierrehumbert, "Phonological Representation: Beyond Abstract Versus Episodic," *Annual Review of Linguistics*, vol. 2, pp. 33–52, 2016.

[14] R. Port and M. O'Dell, "Neutralization of syllable-final voicing in German," *Journal of Phonetics*, vol. 13, pp. 455–471, 1985.

[15] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*, 2018.

[16] D. Byrd, "Articulatory Vowel Lengthening and Coordination at Phrasal Junctures," *Phonetica*, vol. 57, no. 1, pp. 3–16, 2000.

[17] S. Ritter and M. Grice, "The Role of Tonal Onglides in German Nuclear Pitch Accents," *Language and Speech*, vol. 58, no. 1, pp. 114–128, 2015.

[18] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

[19] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology." in *Prosodic Typology: The Phonology of Intonation and Phrasing*, jun, sun-a ed. Oxford: Oxford University Press, 2005, pp. 55–83.

[20] R Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2018. [Online]. Available: http://www.r-project.org/

[21] P.-C. Bürkner, "Advanced {Bayesian} Multilevel Modeling with the {R} Package {brms}," *The R Journal*, 2018.

[22] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software, Articles*, vol. 76, no. 1, pp. 1–32, 2017.

[23] M. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," in *Papers in Laboratory Phonology II: Segment, Gesture, Prosody*, cambridge ed., Cambridge, 1992, pp. 68–86.

[24] K. Iskarous, "The relation between the continuous and the discrete: A note on the first principles of speech dynamics," *Journal of Phonetics*, vol. 64, pp. 8–20, 2017.

[25] A. I. Gafos and S. Benus, "Dynamics of Phonological Cognition," *Cognitive Science*, vol. 30, no. 5, pp. 905–943, 2006.

[26] H. Haken, "Synergetics." Berlin: Springer, 1977.